CLIN 2013

BOOK OF ABSTRACTS OF THE 23RD MEETING OF
COMPUTATIONAL LINGUISTICS IN THE NETHERLANDS:

CLIN 2013

Enschede, January 18, 2013

Theune, M., Nijholt, A., et al.

Book of abstracts of the 23rd Meeting of Computational Linguistics in the Netherlands (CLIN 2013)
Enschede, Universiteit Twente, Faculteit Elektrotechniek, Wiskunde en Informatica

Book orders:
Ms. C. Bijron
University of Twente
Faculty of Electrical Engineering, Mathematics and Computer Science
P.O. Box 217
NL 7500 AE Enschede
tel: +31 53 4893740
fax: +31 53 4893503
Email: C.G.Bijron@utwente.nl

Omslag: Twents Zouthuisje.
Fotografie en ontwerp: Nelson Nijholt

Druk- en bindwerk: Xerox, Enschede.

Preface

Welcome to CLIN 2013, the 23rd meeting of Computational Linguistics in the Netherlands! This year the meeting is organized by the Human Media Interaction group of the University of Twente. After previous meetings in 1994 and 2001, this is the third time the CLIN meeting takes place in Enschede. (See `http://www.let.rug.nl/vannoord/Clin/` for an overview of all previous CLIN meetings.)

Because we like to do things a little bit differently here in Twente, we have chosen a special location for the meeting: Poppodium Atak in the Music Quarter of Enschede. We hope this venue will be a stimulating environment for presenting and discussing new ideas and research directions in computational linguistics and language technology.

More than 90 abstracts were submitted for CLIN 2013, leading to an extensive programme featuring 60 oral presentations and 26 poster/demonstration presentations on a wide variety of topics. Fitting with the tradition of CLIN as having a low threshold for starting researchers, we are pleased to have many presentations by graduate and undergraduate students. At the meeting, we will also present the winner of the STIL Thesis Prize, awarded to the best MA thesis in computational linguistics or its applications.

We are very happy to have as our invited speaker Candy Sidner from Worcester Polytechnic Institute, USA, who will tell us about her current research on conversational agents in her talk "Creating a Real-Time Conversation Manager with multiple time scales". The abstract of her presentation can be found in this booklet.

There will also be a plenary talk by Ineke Schuurman and Marc Kemps-Snijders on the CLARIN pilot "TST Tools voor het Nederlands als Webservices in een Workflow (TTNWW)". Many of our other sponsors have a talk in the Industry Track, which features presentations on language technology research in industry, and is as varied as the rest of the programme.

All authors of accepted abstracts will be invited to submit a full paper to the CLIN journal (see `www.clinjournal.org`). These submissions will undergo a rigorous review process, and a selection of the best papers will appear in the journal.

We thank all our sponsors for their contributions to CLIN 2013. Many thanks are also due to Charlotte Bijron and Alice Vissers for all their help before and during the meeting.

We wish you a very pleasant and inspiring CLIN meeting,

Maral Dadvar
Hendri Hondorp
Anton Nijholt
Mariët Theune
Dolf Trieschnigg
Khiet Truong

Enschede, January 2013

Committees

Program and Organizing Committee

Anton Nijholt	Human Media Interaction, University of Twente, The Netherlands
Mariët Theune	Human Media Interaction, University of Twente, The Netherlands
Dolf Trieschnigg	Human Media Interaction, University of Twente, The Netherlands
Khiet Truong	Human Media Interaction, University of Twente, The Netherlands
Maral Dadvar	Human Media Interaction, University of Twente, The Netherlands
Hendri Hondorp	Human Media Interaction, University of Twente, The Netherlands

Administrative/Technical/Financial Support

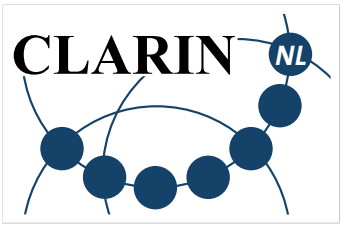
Lynn Packwood	Human Media Interaction, University of Twente, The Netherlands
Charlotte Bijron/Alice Vissers	Human Media Interaction, University of Twente, The Netherlands

CLIN Meetings

October 26, 1990	OTS Utrecht
November 29, 1991	CWI Amsterdam
October 30, 1992	ITK Tilburg
November 25, 1993	Alfa-informatica Groningen
November 23, 1994	Universiteit Twente
December 1, 1995	Antwerpen
November 15, 1996	IPO Eindhoven
December 12, 1997	Letteren KUN, Nijmegen
December 11, 1998	CCL Leuven
December 10, 1999	OTS Utrecht
November 3, 2000	Taal en Informatica, Tilburg
November 30, 2001	Computer Science, University of Twente
November 29, 2002	Alfa-informatica, University of Groningen
December 19, 2003	Centre for Dutch Language and Speech, University of Antwerp
December 17, 2004	Leiden Centre for Linguistics.
December 16, 2005	University of Amsterdam
January 12, 2007	CCL Leuven
December 7, 2007	Radboud University, Nijmegen
January 22, 2009	University of Groningen
February 5, 2010	Utrecht University
February 11, 2011	Hogeschool Gent
January 20, 2012	Tilburg University
January 18, 2013	University of Twente
Winter 2013/2014	INL Leiden
Winter 2014/2015	University of Antwerp

Sponsors

GOLD



[1]

[2]

SILVER

[3]

[4]

BRONZE

[5]

[6]

[7]

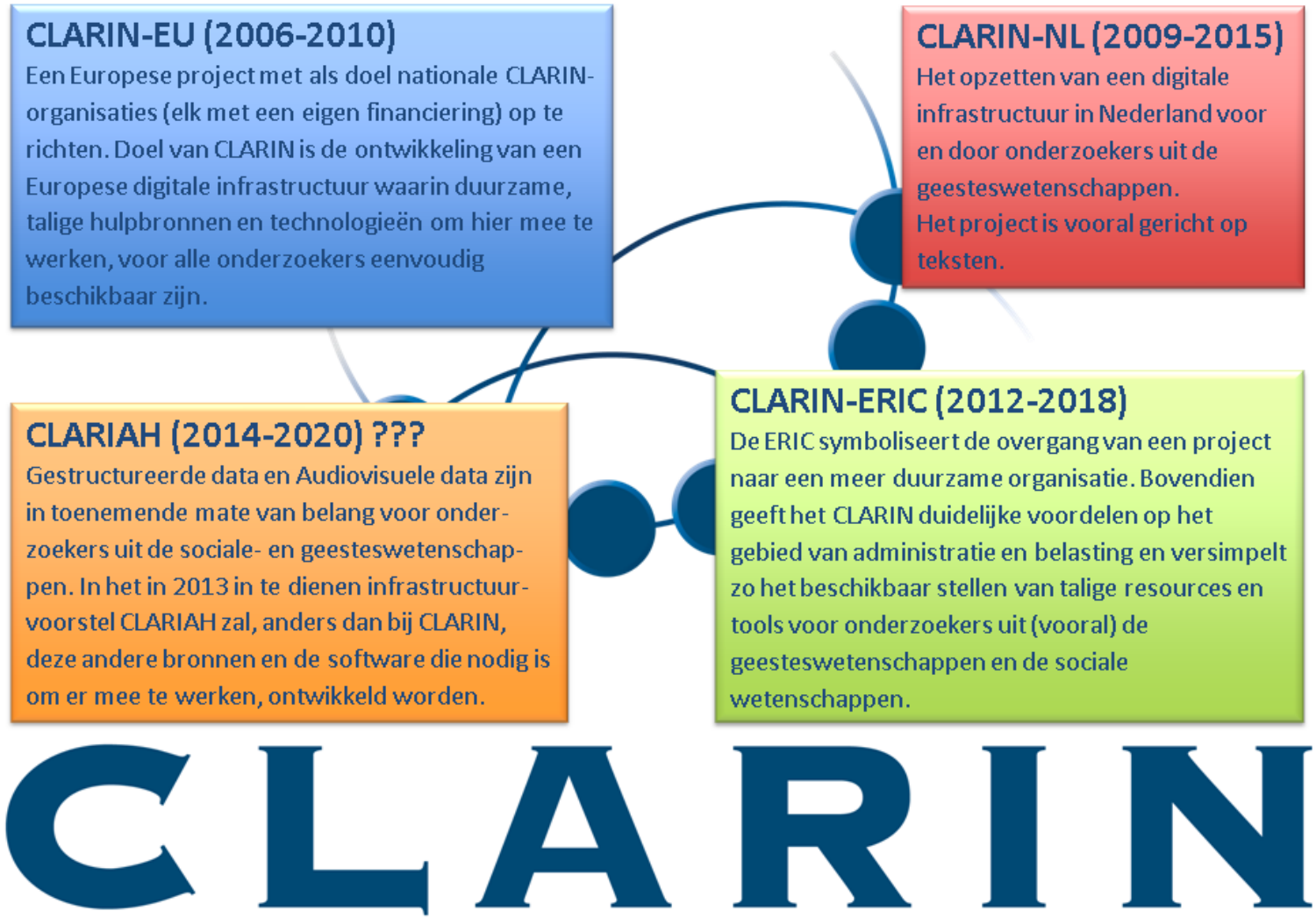
[8]

[9]


[10]

[11]

[1] http://www.clarin.nl
[2] http://www.taalunie.nl
[3] http://www.textkernel.nl
[4] http://www.oracle.nl
[5] http://www.telecats.nl
[6] http://www.notas.nl
[7] http://www.beeldengeluid.nl
[8] http://www.utwente.nl/ctit
[9] http://www.gridline.nl
[10] http://www.informatiewetenschap.org
[11] http://www.siks.nl


CLARIN-EU (2006-2010)
Een Europese project met als doel nationale CLARIN-organisaties (elk met een eigen financiering) op te richten. Doel van CLARIN is de ontwikkeling van een Europese digitale infrastructuur waarin duurzame, talige hulpbronnen en technologieën om hier mee te werken, voor alle onderzoekers eenvoudig beschikbaar zijn.
CLARIN-NL (2009-2015)
Het opzetten van een digitale infrastructuur in Nederland voor en door onderzoekers uit de geesteswetenschappen.
Het project is vooral gericht op teksten.
CLARIAH (2014-2020) ???
Gestructureerde data en Audiovisuele data zijn in toenemende mate van belang voor onderzoekers uit de sociale- en geesteswetenschappen. In het in 2013 in te dienen infrastructuurvoorstel CLARIAH zal, anders dan bij CLARIN, deze andere bronnen en de software die nodig is om er mee te werken, ontwikkeld worden.
CLARIN-ERIC (2012-2018)
De ERIC symboliseert de overgang van een project naar een meer duurzame organisatie. Bovendien geeft het CLARIN duidelijke voordelen op het gebied van administratie en belasting en versimpelt zo het beschikbaar stellen van talige resources en tools voor onderzoekers uit (vooral) de geesteswetenschappen en de sociale wetenschappen.
CLARIN

PROGRAMME

09.00 - 9.30	Registration and coffee (foyer)				
Session 1 9.30-10.50	**Tools (café)**	**Phonology & child language acquisition (kleine zaal)**	**Classification (grote zaal)**	**Text segmentation & text representation (Urenco-zaal)**	**Semantics (theaterfoyer)**
9.30 - 9.50	Novel developments in ELAN (Han Sloetjes, Herman Stehouwer, Sebastian Drude)	Unsupervised pattern discovery in speech (Maarten Versteegh, Michele Gubian, Lou Boves)	Folktale classification using learning to rank (Dong Nguyen, Dolf Trieschnigg, Mariët Theune)	A general-purpose machine learning method for tokenization and sentence boundary detection (Valerio Basile, Johan Bos, Kilian Evang)	Communicative acts and a value-based semantics of business processes (Joris Hulstijn, Rob Christiaanse)
9.50 - 10.10	User friendly signal processing web services (Eric Auer)	Modeling word stress acquisition through the induction of alignment constraints (Jeroen Breteler)	Automatic thematical classification of party programmes (Suzan Verberne, Eva D'hondt, Antal van den Bosch, Maarten Marx)	Automatically identifying compounds (Suzanne Aussems, Sylvie Bruys, Bas Goris, Vincent Lichtenberg, Nanne van Noord, Rick Smetsers, Menno van Zaanen)	Interpreting intentions of speech (Ke Wang, Gerald Penn)
10.10 - 10.30	Facilitating treebank mining of CGN with GrETEL (Liesbeth Augustinus, Vincent Vandeghinste, Frank Van Eynde)	Simulating the acquisition of Dutch word-initial consonant clusters with q-HG (Klaas Seinhorst, Tamás Biró)	Time drift in patent classification (Eva D'hondt, Suzan Verberne, Nelleke Oostdijk, Jean Beney, Kees Koster, Lou Boves)	Learning text representations from character-level data (Grzegorz Chrupala)	Meaning denies structure (Crit Cremers)
10.30 - 10.50	Metadata for tools: a CMDI software profile for describing your software (Eline Westerhout, Jan Odijk)	Linear model for exploring types of vowel harmony (Lili Szabó, Çağrı Çöltekin)	A simple method for topic classification for morpho-logically complex languages (F. Vaassen, J. Kapočiūtė-Dzikienė, G. De Pauw, Walter Daelemans)	A graph-based approach for implicit discourse relations (Yannick Versley)	A tensor-based factorization model of semantic compositionality (Tim Van de Cruys)

10.50 -11.10	Coffee break (foyer)				
11.10-12.45	**Plenary session** (grote zaal)				
11.10 - 11.25	Welcome + STIL thesis prize				
11.25 - 11.45	TST Tools voor het Nederlands als Webservices in Workflow (TTNWW), a CLARIN pilot (Ineke Schuurman, Marc Kemps-Snijders)				
11.45 - 12.45	Invited talk Candy Sidner: Creating a real-time conversation manager with multiple time scales				
12.45 -13.30	Lunch (foyer)				
13.30 - 14.15	**Poster + demo session** (foyer) See the last page of the programme for a list of demonstrations.				
Session 2 14.15-15.35	**Industry Track I (café)**	**Speech, dialogue & assistive interfaces (kleine zaal)**	**Social media(grote zaal)**	**Natural language generation & machine translation (Urenco-zaal)**	**Lexical semantics (theaterfoyer)**
14.15 - 14.35	Automated readability scoring for Dutch texts with SVMs (Roelant Osssewaarde – Edia)	Automatic syllabification using segmental conditional random fields (Kseniya Rogova, Kris Demuynck, Dirk Van Compernolle)	Dealing with big data: the case of Twitter (Erik Tjong Kim Sang, Antal van den Bosch)	Referring expression generation in 3D scenes: modelling human reference production using graphs (Jette Viethen, Emiel Krahmer)	Semantic classification of Dutch and Afrikaans noun-noun compounds (Ben Verhoeven, Walter Daelemans, Gerhard van Huyssteen)
14.35 - 14.55	Controlled automatic translation with Euroglot™ (Leo Konst – Linguistic Systems)	Semantic frame induction in an assistive vocal interface using hierarchical HMMs (Janneke van de Loo, Jort F. Gemmeke, Guy De Pauw, Hugo Van hamme, Walter Daelemans)	Clues for autism in Dutch tweet production (Hans van Halteren, Maarten op de Weegh)	Generation of Dutch referring expressions using the D-TUNA corpus (Marissa Hoek)	Automatic animacy classification for Dutch (Jelke Bloem, Gosse Bouma)

14.55 - 15.15	Why multi-lingual associative networks are better at categorization than their monolingual equivalents (Niels Bloom – Pagelink Interactives)	Toward a model for incremental grounding in dialogue systems (Thomas Visser)	Supporting open-domain event prediction by using cross-domain Twitter messages (Florian A. Kunneman, Ali Hürriyetoglu, Antal van den Bosch)	Incorporating source-side context in machine translation using classifiers (Maarten van Gompel, Antal van den Bosch)	Using an ontology to reduce verb ambiguity (Marten Postma)
15.15 - 15.35	Matching CVs to job vacancies: semantic search in action (Remko Bonnema, Yves Peirsman, Gerard Goossen, Mihai Rotaru, Lena Bayeva, Chao Li, Florence Berbain, Carsten Lygteskov Hansen – Textkernel)	Using natural language as a bridge between pictograph sets (Vincent Vandeghinste)	Event phase identification In textual data (Ali Hürriyetoglu, Florian Kunneman, Antal van den Bosch)	Neural network language models to select the best translation (Maxim Khalilov, Francisco Zamora-Martinez, José A.R. Fonollosa, María José Castro-Bleda, Salvador España-Boquera)	DutchSemCor: in quest of the ideal Dutch semantic corpus (Piek Vossen, Rubén Izquierdo, Attila Görög)
15.35 - 15.55	Coffee break (foyer)				
Session 3 15.55 – 17.15	**Industry track II (café)**	**Spelling correction & normalisation (kleine zaal)**	**Information extraction & web corpora (grote zaal)**	**Machine translation (Urenco-zaal)**	**Lexical semantics, readability & style (theaterfoyer)**
15.55 - 16.15	ZieOok: a recommendation platform for the Netherlands Institute for Sound and Vision (Oele Koornwinder – GridLine)	A “no-no” system (no context, no resource) for classifying a list of words into correctly spelled and misspelled items (Jean-Luc Manguin)	Identification, classification and anonymisation of ‘protected health information’ in real-time medical data for research purposes (Saman Hina, Eric Atwell, Owen Johnson, Claire Brierley)	From old to new Dutch (Sander Wubben, Emiel Krahmer, Antal van den Bosch)	Modelling the acquisition of lexical meaning from caregiver-child interaction: Getting the semantics straight (Barend Beekhuizen , Afsaneh Fazly, Aida Nematzadeh, Suzanne Stevenson)

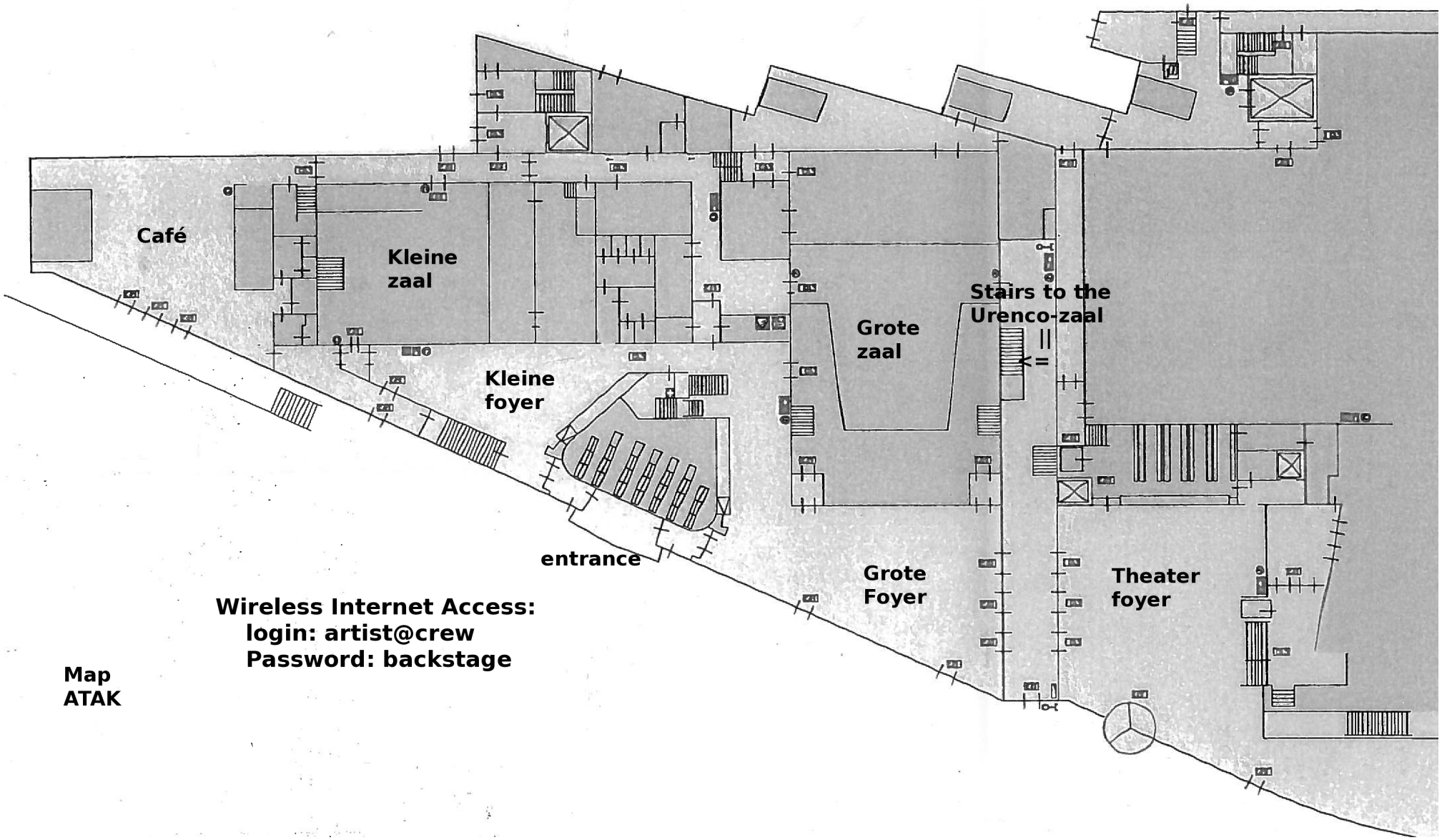
16.15 - 16.35	Virtual assistants: new challenges to language technology (Christian Kissig, Fabrice Nauze, Mandy Schiffrin – Oracle Nederland)	Under the hood of an English context-sensitive spell checker (Wessel Stoop, Antal van den Bosch, Maarten van Gompel, Peter Berck)	An opinion retrieval system for specific domains in the context of the Web 2.0 (Javi Fernández, Carolina Prieto, Elena Lloret, José M. Gómez, Patricio Martínez-Barco, Manuel Palomar)	Disambiguation of word translations without parallel corpora (Erwin Marsi, Björn Gambäck)	Modeling lexical complexity in terms of lexical semantics (Marilisa Amoia)
16.35 - 16.55	Charting the noosphere: mapping internal realities to external worlds (ML Coler – INCAS³)	Building realistic Spellnets for European languages with anagram hashing (Martin Reynaert)	Mass-generation and evaluation of crawler seed URLs for web corpus construction (Adrien Barbaresi, Felix Bildhauer, Roland Schäfer)	On a method of improving the quality of UNL based machine translations (Levon Hakobyan)	Comparing features for automatic readability prediction (Orphée De Clercq, Véronique Hoste)
16.55 - 17.15	How to Search Annotated Text by Strategy? (Roberto Cornacchia, Wouter Alink, Arjen P. de Vries – Spinque)	The Chatty Corpus: a gold standard for Dutch chat normalization (Claudia Peersman, Mike Kestemont, Benny De Decker, Guy De Pauw, Kim Luyckx, Roser Morante, Frederik Vaassen, Janneke van de Loo, Walter Daelemans)	Detecting connected text using an effective unsupervised method (Roland Schäfer)	English to Bangla name transliteration system (Zahurul Islam, Rashedur Rahman)	Cognitive metaphor: externalization of individual blogger (Vera Bakker-Khorikova)
17.15 -19.00	Drinks (café)				

Demonstrations:

Controlled automatic translation with Euroglot™ (Leo Konst)

UBY -- A large-scale unified lexical-semantic resource (Iryna Gurevych et al.)

Interpersonal stance and turn-taking behaviour in a conversational agent system (Merijn Bruijnes et al.)



Wireless Internet Access:
login: artist@crew
Password: backstage

Map
ATAK

Contents

Invited Talk

Oral Presentations

Industrial Track

Posters

Invited Talk

CLIN 2013

Invited talk

Creating a Real-Time Conversation Manager with multiple time scales

Candy Sidner
Computer Science, Fuller Laboratories
Worcester, MA, USA

Abstract:
My current research concerns a conversational agent that is always present in a user's home, can converse and collaborate with a user over an extended period of time, and builds a relationship with the user during that time. To do so we have developed a software architecture that incorporates many types of conversations, and uses real-time sensors and effectors to interact as it converses. The architecture operates with control loops at three distinct time scales, from weeks down to milliseconds. I will motivate the requirements for this architecture, discuss its the novel aspects, and show how it is applied in a relational agent for isolated older adults.

About Candy Sidner:
Candace L. (Candy) Sidner has a long standing interest in human communication and collaboration, and their application to agents, robots, and interfaces, especially those using gesture, social behavior, speech, and natural language.

Candy is a Research Professor at Worcester Polytechnic Institute, with extensive experience in industrial labs. She is principal investigator on the NSF grant Always On Relational Agents for Social Support of Older Adults with Chuck Rich, Worcester Polytechnic Institute (WPI), and Tim Bickmore, Northeastern University. She also serves as an independent consultant on other projects, including the MultiModal Prediction project with Louis-Philippe Morency at ICT, and the Human Robot Interaction project with Chuck Rich at WPI. She has worked on multimodal (gesture) interfaces using a humanoid robot for the role of engagement in conversation, and on interfaces, including those with speech, involving collaborative interface agents in the COLLAGEN project.

She is a Fellow and past Councilor of the American Association for Artificial Intelligence, and a senior member of the IEEE. She serves as an associate editor of the journal Artificial Intelligence, on the editorial boards of the Journal on Multimodal User Interfaces, and ACM Transactions on Interactive Intelligent Systems, and on the scientific advisory boards of IUI, SIGDIAL and HLT-NAACL. She has also been a member of the scientific advisory board for the EU Cognitive Systems for Cognitive Assistants (CoSy) project. She has served as general chair for HLT-NAACL 2007, program cochair of Intelligent User Interfaces 2006, SIGDIAL 2004, chair of Intelligent User Interfaces in 2001, and President of the Association for Computational Linguistics (1989). She received her Ph.D. from MIT in Computer Science.

Abstracts

CLIN 2013

Modeling Lexical Complexity in Terms of Lexical Semantics.

Marilisa Amoia
Department of Applied Linguistics and Machine Translation,
Saarland University, Germany
Email: m.amoia@mx.uni-saarland.de

Automatic text simplification, i.e. the ability of a system to reduce the cognitive complexity of a text and by doing so to render the text more easily understandable by a larger audience, has a wide range of applications which includes applications for the elderly, learners of a second language, children or people with cognitive deficiencies, etc.

Works on text simplification has mostly focused on reducing the syntactic complexity of the text (Siddharthan, 2011; Siddharthan, 2006) and only little work has addressed the issue of lexical simplification (Devlin, 1999; Carroll et al., 1999). The Lexical Simplification Task (Specia et al., 2012) proposed within the SemEval-2012 was the first attempt to explore the nature of the lexical simplification more systematically. Most of the work presented in this context exploit statistical machine learning techniques and target features such as frequency of lexical use, token length, word familiarity. However, (Amoia and Romanelli 2012) attempt at defining lexical complexity in terms of a combination of word frequency and decompositional semantics criteria.

The work described in this paper builds on (Amoia and Romanelli, 2012) and presents the evidence gained by the analysis of the SemEval-2012 Lexical Simplification Task corpus. The original corpus includes lexical items in short contexts that have been manually annotated in terms of cognitive complexity. After applying word sense disambiguation to the lexical items in the corpus, we have studied the impact of lexical semantics in predicting the level of complexity of lexical items.

User Friendly Signal Processing Web Services

Eric Auer
The Language Archive, Max Planck Institute for Psycholinguistics, Nijmegen
Email: eric.auer@mpi.nl

The joint Max Planck Fraunhofer project AVATecH aims to support the very time intensive work of annotating audio and video recordings, letting signal processing modules (recognizers) assist annotators.

We designed a small, flexible framework where XML metadata describes input, output and settings of recognizers. Building blocks are audio and video files, annotation tiers and numerical data, packaged in simple formats. Text pipes allow flexibility in implementation details. The popular TLA ELAN software even lets the user control recognizers directly in their annotation environment: It generates consistent user interfaces for all installed recognizers based on their metadata.

We realized that full recognizers can be inconvenient to install for the user. Hardware, operating system and license requirements can add complexity. AVATecH supported intranet recognizers early, but those are limited by the need for shared network drives between user and server.

Recently, we developed a system where recognizers are run on a server using the free open source CLAM software. With suitable configuration, CLAM can run any command line tool, controlled by remote REST requests. On the user side, only a small proxy tool is installed instead of a real recognizer: The tool dynamically mimicks a recognizer based on the same metadata as before, but actually transfers data to a remote server and back where the real recognizer is installed.

We present details of our setup and workflow, with an outlook towards future extensions within the successor project, AUVIS.

Facilitating Treebank Mining of CGN with GrETEL

Liesbeth Augustinus ✎ *
CCL, KU Leuven
Email: liesbeth@ccl.kuleuven.be

Vincent Vandeghinste
CCL, KU Leuven
Email: vincent@ccl.kuleuven.be

Frank Van Eynde
CCL, KU Leuven
Email: frank@ccl.kuleuven.be

We present the updated version of GrETEL (Greedy Extraction of Trees for Empirical Linguistics, http://nederbooms.ccl.kuleuven.be/eng/gretel), a search engine for querying treebanks by example. As input, GrETEL requires a natural language example, containing the linguistic construction one is looking for. After indicating the relevant information, the input example is automatically parsed and converted into an XPath expression, which is used to look for similar sentences in the treebank. GrETEL makes it thus possible to query treebanks without knowledge of (complicated) formal query languages or data formats.

In the first version of GrETEL, only written Dutch (LASSY) was included. Based on user requests we have now included the Spoken Dutch Corpus (CGN) as well.

Building XPath expressions for LASSY is done using the POS tags (e.g. 'noun', 'verb') assigned by the ALPINO parser. CGN does not contain those tags; only the CGN/D-COI tags (e.g. 'N(soort,ev,basis,onz,stan)', 'WW(inf,vrij,zonder)') are included. In order to generate formal queries for CGN, we have added a POS tagger to the system. After POS tagging the input sentence, GrETEL adds the CGN/D-COI tags to the ALPINO parse, which is then converted into an XPath expression.

The adaptations that were necessary to include CGN into GrETEL are also used to extend and improve the search engine for LASSY. As the CGN/D-COI tags contain more detailed information compared to the more general ALPINO tags, it is possible to use this extra information to refine the (automatically generated) query.

*Smallpencil ✎: means corresponding author

Automatically Identifying Compounds

Suzanne Aussems
Tilburg University, Tilburg, The Netherlands
Email: s.h.j.a.aussems@uvt.nl

Sylvie Bruys
Tilburg University, Tilburg, The Netherlands
Email: s.bruys@uvt.nl

Bas Goris
Tilburg University, Tilburg, The Netherlands
Email: b.c.c.goris@uvt.nl

Vincent Lichtenberg
Tilburg University, Tilburg, The Netherlands
Email: v.j.j.lichtenberg@uvt.nl

Nanne van Noord
Tilburg University, Tilburg, The Netherlands
Email: n.j.e.vannoord@uvt.nl

Rick Smetsers
Tilburg University, Tilburg, The Netherlands
Email: r.h.a.m.smetsers@uvt.nl

Menno van Zaanen ✎
Tilburg University, Tilburg, The Netherlands
Email: mvzaanen@uvt.nl

The AuCoPro project deals with the analysis of compounds in two closely-related languages: Afrikaans and Dutch. The underlying idea is that comparing aspects of closely-related languages will lead to new knowledge about these languages and their similarities and differences. In particular, in the AuCoPro project, new resources are developed for the segmentation and semantic analysis of compounds in both Afrikaans and Dutch. However, before the analysis of compounds can be performed, having access to a list of compounds is essential.

In this context, we present the development of a system that aims to identify compounds. Given that we require compounds in both Dutch and Afrikaans, we aim to make as few language dependent choices as required and we keep track of such choices made during the development of the system. The system, which is currently still being developed further, is based on both symbolic (searching for known words as parts of a compound) as well as machine learning approaches.

The evaluation is currently performed on Dutch text. Taking the word frequency information from the SoNaR corpus as the starting point, we cleaned up the list, removing obviously non-Dutch words. Part of the remaining words are manually annotated and used for evaluation purposes. Additionally, the system is evaluated on the e-lex corpus.

Cognitive Metaphor: Externalization of Individual Blogger

Vera Bakker-Khorikova
Moscow State University, Russia
Email: vera.khorikova@gmail.com

The paper tackles the challenging matter of linguistic behaviour of the modern English-speaking online bloggers taking into consideration their creativity and individuality. Thanks to the inborn human ability to express abstract matters in a linguistic form, online bloggers successfully apply the powerful tool of cognitive metaphor to externalize ideas, thoughts and dreams unconsciously. The research is based on writing of various individual bloggers, seen as a possibility to express own thoughts based on individual experience and perception of the surrounding world.

The major focus was placed on the role of cognitive metaphor traced in individual blogs as well as its frequency and models. Systematization through comparative analysis of metaphorical models used by online bloggers as well as parallels to peculiarities of the bloggers' cognition and psychological traits have resulted in significant results to postulate that the inner world of the online blogger can be observed by analysis of the metaphorical expressions used in individual blogs.

Based on the recent research in the field of psycholinguistics and pragmatics, the study applies these findings to online discourse, as well as draws parallels with socio-cultural, psychological and pragmatic aspects of online behaviour of bloggers. The research aims at tackling the challenging subject of online discourse in combination with socio-cultural, cognitive background of the users of online journals (blogs) as well as exposure of their personality to the readers of blogs.

Mass-Generation and Evaluation of Crawler Seed URLs for Web Corpus Construction

Adrien Barbaresi
German Language and Linguistics, FU Berlin
Email: adrien.barbaresi@fu-berlin.de

Felix Bildhauer
Sonderforschungsbereich 632/A6, FU Berlin
Email: felix.bildhauer@fu-berlin.de

Roland Schäfer ✎
German Language and Linguistics, FU Berlin
Email: roland.schaefer@fu-berlin.de

We describe an infrastructure for harvesting crawler seed URLs for diverse languages specifically for the purpose of web corpus construction. Since search engines no longer provide free API access to obtain lots of URLs (making the BootCaT method obsolete), we acquire URLs by harvesting social messaging services, Wikipedia, dmoz.org, and by steadily querying meta-search engines with random word tuples using long politeness intervals. We classify the language of the documents behind the URLs using langid.py. The resulting database contains URLs of documents classified by language, ranging from several millions for English to a few hundred thousands for Indonesian and Dutch. The database will be publicly accessible. We evaluate the quality of the URLs in terms of the amounts of linguistically relevant data gathered by using the seeds. We built test corpora by (1) fetching 1,000 URLs directly from the database per language and URL source, (2) doing breadth-first crawls using these URLs as seeds. Among other things, we compare the amount of data left over by diverse filters (including perfect/near-duplicate and connected-text detection) in the texrex toolchain (Schäfer & Bildhauer, 2012) for all test corpora. The amount of documents left over from a crawl correlates strongly with the amount left over in the seed-only corpora (r=0.89). Consequently, we will process all URLs in the database with texrex and generate meta-information encoding their usefulness for crawling/corpus construction.

Schäfer, R. & F. Bildhauer (2012) Building large corpora from the web using a new effcient tool chain. in: Proceedings LREC'12, Istanbul, ELRA, 486-493.

A General-Purpose Machine Learning Method for Tokenization and Sentence Boundary Detection

Valerio Basile ✎
University of Groningen
Email: v.basile@rug.nl

Johan Bos
University of Groningen
Email: johan.bos@rug.nl

Kilian Evang
University of Groningen
Email: k.evang@rug.nl

Tokenization is widely regarded as a solved problem due to the high accuracy that simple rule-based tokenizers achieve. Many of these rules must consider language-specific punctuation rules as well as domain-specific abbreviations and notation conventions. One problem with this is that adaptating a rule-based tokenizer to a new language or domain requires explicit changing and testing of the rules.

We show that high-accuracy tokenization can be obtained by using supervised machine learning methods that require training on only a relatively small amount of tokenized texts from the target language and domain. Tagging is performed on the character level, following an IOB tagging scheme that has been adapted for simultaneous tokenisation and sentence boundary detection.

Every character is labbeled as either I, O, T, or S. The first character of each token is labeled as T, or if it is the first character of a sentence, as S. Other characters are labeled as I (part of a token) or O (not part of a token). This is a flexible scheme, allowing us, for instance, to skip hyphens at line breaks as in "ergo-nomic", or to split tokens in contractions such as "don't".

To adapt the tokenizer to a new domain, it suffices to retrain it on appropriately tokenized data without any need for explicit changing of manually-coded rules. We show the effectiveness of our method by tokenizing open-domain text in English and Dutch, as well as biomedical text used in the BioCreative task 1A.

Modelling the Acquisition of Lexical Meaning from Caregiver-child Interaction: Getting the Semantics Straight

Barend Beekhuizen ✎
Leiden University
Email: barendbeekhuizen@gmail.com

Afsaneh Fazly
University of Toronto
Email: afsaneh@cs.toronto.edu

Aida Nematzadeh
University of Toronto
Email: aida@cs.toronto.edu

Suzanne Stevenson
University of Toronto
Email: suzanne@cs.toronto.edu

With a few recent exceptions, computational models of the acquisition of lexical meaning rarely employ information about the actual objects and events present in the situational context to learn the meaning from. Rather, synthesized meaning on the basis of the language is often used (i.e., if a verb grab is uttered, the meaning 'grab is assumed to be present too). This approach coarsely underestimates the noise and referential uncertainty found in actual situational contexts of caregiver-child interaction: many events do not take place when the words denoting them are used and the words referring to an object or event are most often not present when the object or event is.

In this talk, we show how a well-studied word learning model (Fazly et al. 2010), that performs well on synthetic semantic data, does not scale well when the input data consists of child-directed speech with the objects and events from the actual situation. We discuss the likely sources of this problem, show that these are not due to the particular model, and present several cognitively realistic, simple extensions of the model that improve the performance. The research shows that understanding the mechanisms of language acquisition through computational modelling depends to a large extent on the naturalism of the input data.

Fazly, A., Alishahi, A. & Stevenson, S. (2010). A Probabilistic Computational Model of Cross-situational Word Learning, Cognitive science 34, 1017-1063.

Automatic Animacy Classification for Dutch

Jelke Bloem ✎
University of Groningen
Email: j.bloem.3@student.rug.nl

Gosse Bouma
University of Groningen
Email: g.bouma@rug.nl

We present an automatic animacy classifier for Dutch that can determine the animacy status of nouns - how alive the noun's referent is (human, inanimate, etc.). Animacy is a semantic property that has been shown to play a role in human sentence processing, felicity and grammaticality ("the spoon *who is on the table fell."). We expect knowledge about animacy to be helpful for parsing, translation and other NLP tasks, although animacy is not marked explicitly in Dutch.

Only a few animacy classifiers and animacy-annotated corpora exist internationally. For Dutch, animacy information is only available in the Cornetto lexical-semantic database. We augment this lexical information with context information from the Dutch Lassy Large treebank, to create training data for an animacy classifier that uses context features.

An existing Swedish animacy classifier (Øvrelid, 2009) uses the k-nearest neighbour algorithm with morphosyntactic distributional features, e.g. how frequently the noun occurs as a sentence subject in a corpus, to decide on the (predominant) animacy class. For Dutch we use the same algorithm, but with distributional lexical features, e.g. how frequently the noun occurs as a subject of the verb 'to think' in a corpus. The size of the Lassy Large corpus makes this possible, and the higher level of detail these word association features provide, increases the classifier accuracy and provides us with accurate Dutch-language animacy classification. These results allow (semi-)automatic corpus animacy annotation for creating animacy training resources, which can help other Dutch NLP tools to incorporate the animacy property of nouns.

From Old to New Dutch

Antal van den Bosch
Radboud University Nijmegen
Email: a.vandenbosch@let.ru.nl

Sander Wubben ✎
Tilburg University
Email: s.wubben@uvt.nl

Emiel Krahmer
Tilburg University
Email: e.j.krahmer@uvt.nl

In this talk we will discuss a method to translate between diachronically distinct language variants. For research into history, historical linguistics and diachronic language change, historical texts are of great value. Specifically from earlier periods, texts are often the only forms of information that have been preserved. One problem that arises when studying these texts is the difference between the language the text is written in and the modern variant that the researchers who want to study the texts know and speak themselves. It takes a great deal of deciphering and interpretation to be able to grasp these texts. Our aim is to facilitate laymen in studying medieval texts by attempting to generate literal translations of the sentences in the text into modern language. In particular we focus on the task of translating Middle Dutch to modern Dutch.

Modeling Word Stress Acquisition through the Induction of Alignment Constraints

Jeroen Breteler
Utrecht University
Email: j.m.w.breteler@students.uu.nl

Previous literature on the computational modeling of word stress acquisition stipulates the constraints that must be ranked to make up the final grammar (Apoussidou 2006). In contrast, the present proposal models acquisition through constraint induction. Using words as input, the model makes statistical observations on the alignment of prosodic categories. Hence, it is online and frequency-dependent, similar to GLA (Boersma 1997). Consistent alignment behavior is generalized into constraints (cf. StaGe, Adriaans & Kager 2010). In this way, the model builds an OT grammar that reflects the metrical stress system (Hayes 1980) of the input language. The model drives on the intuition that the strongest constraints are most salient in the input data, so there is no need for a reranking mechanism (Smolensky & Tesar 2000). Crucially, the model does not receive information on the prosodic structure of input words; it uses the induced constraints to reduce parsing ambiguity. The model allows for different specifications of the prosodic hierarchy, facilitating comparisons between more traditional views and new proposals like Weakly Layered Feet (Kager 2012). Results show excellent coverage for single-type stress languages with binary rhythms. Testing on languages with a primary/secondary stress contrast and languages with ternary rhythms is still in progress.

Learning Text Representations from Character-level Data

Grzegorz Chrupała
Saarland University / Tilburg University
Email: gchrupala@lsv.uni-saarland.de

Automatically learned word representations have been used in much recent work in computational linguistics. They have proved helpful as high-level features and have contributed to state-of-the-art results in information extraction, parsing and other areas (Turian et al. 2010, Chrupała 2011). However, assuming sequences of word tokens as input to linguistic analysis is often unjustified. For many languages word segmentation is a non-trivial task. Naturally occurring text is often a complex mix of natural language strings intermingled with other character data.

In this paper we propose to learn text representations directly from raw character sequences. We train a recurrent neural network to predict the next character in a large amount of text by truncated backpropagation through time (Mikolov et al. 2010). The network learns the task by generalizing over character sequences and using its hidden layer to evolve the abstract representations of the character sequences it sees.

We demonstrate that the learned text representation are useful by focusing on a practical character-level text analysis task. We use HTML markup in a large collection of questions posted to a programming Questions and Answers forum as a supervision signal in order to learn to recognize spans of text containing programming language code. We use a supervised Conditional Random Field model with n-gram features as a baseline learner. By enriching the feature set with the text representations learned from raw character data by the recurrent neural network we are able to achieve substantially better performance on this task.

Comparing Features for Automatic Readability Prediction

Orphée De Clercq ✎
LT3, University College Ghent, Belgium
Email: orphee.declercq@hogent.be

Véronique Hoste
LT3, University College Ghent, Belgium
Email: veronique.hoste@hogent.be

What is it that makes a particular text easy or difficult to read? This question has been central to the HENDI project which now reaches an end. Within this project a corpus of generic Dutch and English texts has been assessed on readability by two user groups (experts and crowd). Based on these assessments and previous research on readability prediction, we were able to derive various feature groups that might be good indicators of text readability.

In this talk we focus on the Dutch assessments; the texts were derived from the semantically and syntactically annotated SoNaR1 corpus which makes it a particularly interesting data set. It allowed us to perform a qualitative analysis on the following feature groups: classical (based on more traditional metrics), lexical, syntactic and discourse features. This was done by comparing readability predictions based on gold standard information with those based on automatically-derived features. This is to our knowledge the first study, especially for the semantic and discourse features, envisaging to grasp the upper bound of automatic readability prediction. As machine learning methods for these automatic predictions, we use linear regression, multiclass and binary classification.

Meaning Denies Structure

Crit Cremers
LUCL / Universiteit Leiden
Email: c.l.j.m.cremers@hum.leidenuniv.nl

Referring to the construction of a grammar- and meaning-driven parser and generator, I will defend the following conjecture: if a grammar is a formal system, dealing with constituents and entailments, and if constituents and entailments are empirical objects, a functional relation between constituents and entailments cannot exist. In particular, I will argue that theorems of the form 'this structured sentence entails that structured sentence' can neither be verified by unification of signs nor by composition of meanings. Constituents and entailments live by different algebras.This state of affairs reflects the nature of the grammar's incompleteness as a formal system. As such, this incompleteness is an asset, rather than a weakness of formal grammar - of every grammatical model that takes both syntax and semantics serious. As a consequence, a functional route from form to meaning or back is not available in serious grammar. Either we stick with underspecification of meaning, giving up full interpretation, or we have to leave the idea that real interpretation can be modelled compositionally. Good syntax cannot buy us deep interpretation - nor the other way around - no matter whether the syntax is Bayesian or symbolic. At the end, we need (lexical) oracles to get the semantics right, since serious interpretation denies valid structure. These lexical receipts are complex reflections of structure and composition. An exemplary lexical oracle providing the semantics of exception-sentences will be presented.

A Tensor-based Factorization Model of Semantic Compositionality

Tim Van de Cruys
IRIT & CNRS - Toulouse
Email: tim.vandecruys@irit.fr

In this presentation, a novel method for the computation of compositionality within a distributional framework is presented. The key idea is that compositionality is modeled as a multi-way interaction between latent factors, which are automatically constructed from corpus data. The method is used for the composition of subject-verb-object combinations. First, a latent factor model for both nouns and verbs is computed from standard co-occurrence data. Next, the latent factors are used to induce a latent model of three-way subject-verb-object interactions. The model has been evaluated on a similarity task for transitive phrases, in which it exceeds the state of the art.

Time Drift in Patent Classification

Eva D'hondt ✎
Radboud Universiteit Nijmegen
Email: e.dhondt@let.ru.nl

Suzan Verberne
Radboud Universiteit Nijmegen
Email: s.verberne@cs.ru.nl

Nelleke Oostdijk
Radboud Universiteit Nijmegen
Email: n.oostdijk@let.ru.nl

Jean Beney
Insa de Lyon
Email: jean.beney@ouvaton.org

Kees Koster
Radboud Universiteit Nijmegen
Email: kees@cs.ru.nl

Lou Boves
Radboud Universiteit Nijmegen
Email: l.boves@let.ru.nl

For many text classification tasks where the data is collected over a period of time (years or even decades), its underlying distribution is likely to change over time. A prime example of this 'time drift' in large corpora can be found in patent classification. For example, given the explosion of new mobile technologies in the last decade, relatively little overlap can be expected between Electric Communication Devices patents granted in the 70's and the patent applications that are filed today, both in terms of terminology and the concepts described. Most of the existing research on improving automated patent classification ignores this temporal dimension of the data and treats their training corpus as static but noisy. However, in the patent classification practice, incoming patent applications have to be classified with a classification system trained on available (and therefore older) patents. In this presentation, we will show the existence of time drift in the CLEF-IP '11 corpus, a patent corpus that is commonly used for patent classification experiments, and discuss the impact it has on the classification accuracy of the most recent patent applications. Furthermore, we will discuss how we can improve classification accuracy of recent applications through careful term and document selection in the training data. Preliminary results have shown the existence of time drift in this corpus, and the temporal mismatch between training and test material. In the selection of training data, we found a way to optimize the trade-off between sample size and the recency effect.

An Opinion Retrieval System for Specific Domains in the Context of the Web 2.0

Javi Fernández ✎
Universidad de Alicante
Email: javifm@ua.es

Carolina Prieto
AITEX, Instituto Tecnológico Textil
Email: cprieto@aitex.es

Elena Lloret
Universidad de Alicante
Email: elloret@dlsi.ua.es

José M. Gómez
Universidad de Alicante
Email: jmgomez@ua.es

Patricio Martínez-Barco
Universidad de Alicante
Email: patricio@dlsi.ua.es

Manuel Palomar
Universidad de Alicante
Email: mpalomar@dlsi.ua.es

In chemical textile domain, experts have to analyse chemical components and substances that might be harmful for their usage in clothing and textiles. Therefore, it is crucial for experts to have access to the information people is expressing on the Internet, to be able to warn the potential risks of using specific substances. This could be achieved by automatically analysing the information available on the Social Web, searching for comments, reports, opinions, that people have expressed concerning specific products. For this reason, the objective of this study is to design, implement and evaluate an opinion retrieval system able to detect and classify opinions related to the chemical textile domain that appear in the Web 2.0. This is a very challenging domain because there is a small amount of opinionated information available and the existing opinions are unbalanced: the number of complaints is higher than the positive opinions. We describe the creation of several sentiment datasets with documents from this domain containing forbidden substances and perform a series of experiments using machine learning techniques and semantic resources, such as WordNet. Despite the small size of the datasets created and the challenges mentioned, our approach is comparable to the state-of-the-art systems with an F-score of 65% for sentiment polarity (distinguishing between positive, negative, and neutral), and 82% for negativity (distinguishing only between negative and non-negative sentences). In both cases, the results are very promising, thus encouraging to continue working on the improvement of the proposed system.

Incorporating Source-side Context in Machine Translation using Classifiers

Maarten van Gompel ✎
Radboud University Nijmegen
Email: proycon@anaproy.nl

Antal van den Bosch
Radboud University Nijmegen
Email: a.vandenbosch@let.ru.nl

Statistical Machine Translation can benefit from source-language side context information. Traditional phrase-based SMT systems translate phrases without regard to this source-side context, whereas information about the target-language context plays its part through the language model component. Source-side context information can however be incorporated by integrating machine learning classifiers that map phrases given their context to their translations. Source-side context needs not consist of only local context features, but also global context keywords play a promising role. We explore and compare various classifier configurations; including construction experts. These are classifiers that are trained to be experts in translating a single phrase, given various contexts. In doing so we bring proven techniques from Word Sense Disambiguation to Machine Translation. Translation quality gains are expected when state-of-the-art SMT systems employing a stack-based decoding algorithm are enriched in such a fashion. A novel decoder that directly integrates classifier support has been developed to this end.

Event Phase Identification In Textual Data

Ali Hürriyetoğlu ✎
Radboud University
Email: a.hurriyetoglu@let.ru.nl

Florian Kunneman
Radboud University
Email: f.kunneman@let.ru.nl

Antal van den Bosch
Radboud University
Email: a.vandenbosch@let.ru.nl

Events are mainly described in textual data by domain terms, verbs, time expressions, place names and participant information. Human readers understand features and the phase of the event by decoding these signals. Textual descriptions of events change with the time of the event and with the time the event is described. Therefore, analysis of this change in linguistic structure may provide insights to be implemented in information systems in order to identify an event phase automatically.

Event phase signals are considered to include but are not limited to time expressions, domain terms, verb forms, and linguistic structures related to intention and epistemics. Information systems can use these signals to predict and understand current events and analyze historical events.

Although changing patterns of event phase signals are likely to differ among types of event (social, official, short term, long term), events always have some begin, some 'during' phase, and some end. Our analysis focuses on general categories of linguistic structures that play a role in distinguishing between generic event phases. Understanding this pattern can provide us a language- and domain-independent topic detection methodology.

Our study describes the importance and challenge of identifying event phase signals and their patterns of change in different phases of a particular event. The initial results will be presented for three main event phases, before, during and after an event. A corpus of Dutch microtexts (tweets) is used to extract signal patterns.

On a Method of Improving the Quality of UNL Based Machine Translations

Levon Hakobyan
Russian-Armenian (Slavonic) University, Armenia, Yerevan
Email: levon.r.hakobyan@gmail.com

A method of improving UNL based machine translations quality using predefined samples is introduced. UNL is an artificial language which is being used as an interlingua during machine translations. We claim that one of the most appropriate approaches for improving the quality of UNL based machine translations is to let machine to generate translation rules itself using predefined samples. Sample is a pair that consists of one sentence on UNL with its translation on natural language made by specialists. To use this approach there is a need to have a special function to compare natural language sentences. This function should be used to get the 'distance' between sentence from sample and sentence obtained by translation in order to estimate the quality of translation. We also introduce first results of proposed method implementation.

Clues for Autism in Dutch Tweet Production

Hans van Halteren ✎
Radboud University Nijmegen
Email: hvh@let.ru.nl

Maarten op de Weegh
Autimaat
Email: M.opdeWeegh@autimaat.nl

It is well known that people who have been diagnosed as having a disorder in the autistic spectrum also show patterns of communication and language use that differ from the general norm. This would imply that these people should be identifiable by their writings, using author profiling software. In this paper we test whether this is true by comparing samples of the tweet production by Dutch authors who are known to have been diagnosed in the autistic spectrum to samples of the tweet production of a random selection of Dutch Twitter users. We selected our positive test subjects either on the basis of personal knowledge or on the basis of Twitter profile information, leading to a group of 26 positive authors, who produced at least 1000 words on Twitter in 2011. The control group was formed by a selection of 1158 authors who were equally active in the same period. For author profiling, we used the Linguistic Profiling system (van Halteren, 2004), with uni-, bi- and trigrams of tokens as features. In the evaluation, we selected optimal hyperparameters on the basis of a development set of 11 positive subjects. Of the remaining 15 positive subjects, only 1 was rejected by the profile (7%), while also 7% of the control group was (probably) falsely accepted. Subsequently, we examined the features that played a large role in the recognition process and found that a good number of these could be linked to communication patterns described in the literature on autism.

Identification, Classification and Anonymisation of 'Protected Health Information' in Real-time Medical Data for Research Purposes

Saman Hina ✎
University of Leeds, NED University of Engineering and Technology
Email: scsh@leeds.ac.uk

Eric Atwell
University of Leeds
Email: E.S.Atwell@leeds.ac.uk

Owen Johnson
University of Leeds
Email: O.A.Johnson@leeds.ac.uk

Claire Brierley
University of Leeds
Email: C.Brierley@leeds.ac.uk

Protection of information that identifies an individual is an important issue in the medical domain, where the research community is encouraged to use real-time data sets for research purposes. These data sets contain both structured data-fields (e.g. Name, Age) and unstructured narrative annotations that can be used by researchers in various disciplines including computational linguistics. However, these real-time data sets cannot be distributed without anonymisation of 'Protected Health Information'; PHI is information such as name, age, etc. that can identify an individual. It is easy to remove PHI in structured data-fields but it is much harder to eliminate PHI from the narratives. Therefore, we present an anonymisation system using a challenging corpus containing medical narratives and medical codes. The corpus used in this research contains 2534 PHIs in 1984 medical records. 15% of the labelled corpus was used for the development and 85% was held out for the evaluation. Our anonymisation system follows a two-step process; 1) Identification and classification of PHIs with four PHI categories ('Patients Name', 'Doctors Name', 'Other Name [Names excluding patients and doctors]', 'Place Name'), 2) Anonymisation of PHIs by replacing identified PHIs with Tags corresponding to their respective PHI categories. We adopted rule-based approach to identify, classify and anonymise PHIs with PHI categories. As a baseline, we modified the standard GATE named-entity recogniser to detect the PHIs; this scored 76% F-measure. Our system outperformed GATE NER application on human-annotated gold standard, with 100% F-measure. This demonstrated the reliability of our approach to this research task.

Generation of Dutch Referring Expressions using the D-TUNA Corpus

Marissa Hoek
University of Twente
Email: m.d.hoek@student.utwente.nl

The topic of my research is generating referring expressions in Dutch: i.e. noun phrases that serve as descriptions of specific objects or people. The generation of referring expressions is usually done in two steps: attribute selection and realisation in natural language. I focus only on the realisation step: generating a noun phrase from given attributes. My research is done on the Dutch version of the TUNA-corpus, which contains annotated human-generated descriptions of furniture objects and people.

I developed four algorithms for the realisation task, each an improvement over the last. The first version was a directly translated version of an English system developed by Ivo Brugman (2009). The second version was an improved baseline which used simple Dutch grammatical rules. The third algorithm used the result of a manual corpus analysis to use the words and phrases most often seen in the corpus. The final version used templates automatically generated from the corpus, which specified the order of attributes.

I then evaluated the algorithms using a program which tested for string similarity, and a human evaluation which tested clarity and fluency. I observed a steady improvement of scores in both the automatic and human evaluation for each new version of the algorithm. I used a method to extract the original attribute choices from the corpus, and the results were compared to automatically generated attribute selection using the GRAPH-algorithm. Using the original attributes had a positive effect on string similarity, but no substantial difference on the clarity and fluency scores.

Communicative Acts and a Value-based Semantics of Business Processes

Joris Hulstijn ✎
Delft University of Technology
Email: j.hulstijn@tudelft.nl

Rob Christiaanse
Delft University of Technology and EFCO Solutions
Email: r.christiaanse@efco-solutions.nl

Business communication is increasingly being formalized. Business processes consist of activities, many of which have the character of a communicative act (Winograd 1987). The semantics of these communicative acts can often be characterized in terms of commitment change (Singh 1998). Accepting a purchase order constitutes or - counts as - a commitment to deliver, compare (Searle 1995). Traditionally, accounting systems have been used to record both financial transactions and commitments. For different types of business, there are normative models of the flow of money and goods (Starreveld et al. 1994). These models provide reconciliation relations: equations that should hold for all transactions (e.g. debit = credit; begin + in - out = end).

In this paper we argue that exchanges of economic value - as laid down in an accounting information system - can provide a formal semantics to business processes. Such a semantics can be used to derive invariants and constraints for the sound and complete specification of business processes. The concepts are illustrated by an automated control system used to monitor compliance to contracts regarding healthcare related public transport services.

References

Searle, J. R. (1995) The Construction of Social Reality, The Free Press.
Singh, M. P. (1998) Agent communication languages: Rethinking the principles, IEEE Computer, 31(12), 40-47.
Starreveld, R. W., de Mare, B. and Joels, E. (1994) Bestuurlijke Informatieverzorging (in Dutch), Samsom, Alphen aan den Rijn.
Winograd, T. (1987) A Language/Action Perspective on the Design of Cooperative Work, Human-Computer Interaction, 3(1), 3-30.

English to Bangla Name Transliteration System

Zahurul Islam ✎
AG Texttechnology , Goethe University Frankfurt
Email: zahurul@em.uni-frankfurt.de

Rashedur Rahman
AG Texttechnology , Goethe University Frankfurt
Email: kamol.sustcse@gmail.com

Machine translation systems always struggle transliterating names and unknown words during the translation process. It becomes more problematic when the source and the target language use different scripts for writing. To handle this problem, transliteration systems are becoming popular as additional modules of the MT systems. In this abstract, we are presenting an English to Bangla name transliteration system that outperforms Google's transliteration system. The transliteration system is the same as the phrase based statistical machine translation system, but it works on character level rather than on phrase level.

The performance of a statistical system is directly correlated with the size of the training corpus. In this work, 2200 names are extracted from the Wikipedia cross lingual links and from Geonames . Also 3694 names are manually transliterated and added to the data. 4716 names are used for training, 590 for tuning and 588 names are used for testing.

If we consider only the candidate transliterations, the system gives 64.28% accuracy. The performance increases to more than 90%, if we consider only the top 5 transliterations. To compare with the Google's English to Bangla transliteration system, a list of 100 names are randomly selected from the test data and translated by both systems. Our system gives 63% accuracy where the Google's transliteration system does not transliterate a single name correctly. We have found significant improvement in terms of BLUE and TER score when we add the transliteration module with an English to Bangla machine transliteration system.

Neural Network Language Models to Select the Best Translation

Maxim Khalilov ✎
Centre de Recerca TALP,
Universitat Politècnica de Catalunya
Email: maxim@translationautomation.com

Francisco Zamora-Martinez
Departament de Ciencias Físicas, Matemáticas y de la Computación,
Universidad CEU Cardenal Herrera
Email: francisco.zamora@uch.ceu.es

José A.R. Fonollosa
Centre de Recerca TALP,
Universitat Politècnica de Catalunya
Email: adrian@gps.tsc.upv.edu

María José Castro-Bleda
Departamento de Sistemas Informáticos y Computación,
Universitat Politècnica de València
Email: mcastro@dsic.upv.es

Salvador España-Boquera
Departamento de Sistemas Informáticos y Computación,
Universitat Politècnica de València
Email: sespana@dsic.upv.es

The quality of translations produced by statistical machine translation (SMT) systems crucially depends on the generalization ability provided by the statistical models involved in the process. While most modern SMT systems use N-gram models to predict the next element in a sequence of tokens, we follow a continuous space language model (LM) based on neural networks (NN). Experimental results on a small Italian-to-English and a large Arabic-to English translation tasks, which take into account different word history lengths (N-gram order), show that the NNLMs are scalable to small and large data and can improve an N-gram-based SMT system.

A considerable improvement has been obtained using a NNLM for Italian-to-English translation. The best-performed 4-gram NNLM system allows a gain up to 1.2 BLEU points for the test set over the system that includes conventional N-gram LM as a feature in the decoder.

For Arabic-to-English translation, the system configuration providing the better BLEU score corresponds to the 5-gram LMs. Incorporating NNLMs into this N-gram-based SMT system allows gaining up to 0.7 BLEU points for the test set over the baseline. Increase of N-gram order to 6 does not lead to further performance improvement.

Supporting Open-Domain Event Prediction by Using Cross-Domain Twitter Messages

Florian A. Kunneman ✎
Centre for Language Studies, Radboud University,
P.O.Box 9103, 6500 HD Nijmegen, The Netherlands
Email: f.kunneman@let.ru.nl

Ali Hürriyetoğlu
Centre for Language Studies, Radboud University,
P.O.Box 9103, 6500 HD Nijmegen, The Netherlands
Email: a.hurriyetoglu@let.ru.nl

Antal van den Bosch
Centre for Language Studies, Radboud University,
P.O.Box 9103, 6500 HD Nijmegen, The Netherlands
Email: a.vandenbosch@let.ru.nl

While much information is conveyed in the continuous stream of social media messages, it is hard to signal significant emergent processes that are less prominent or frequent than trending topics, but have the potential of becoming high-impact events. Magnifying these and accordingly predicting (social) news events can be of great assistance to journalists who want to be as informed as possible.

In order to identify emerging events, a linguistic *pattern* needs to be grasped within and among messages, denoting a collective future action, anxiety or concern: a pattern of anticipation. While this task asks for an open-domain approach, corresponding to the goal of detecting future *unknown* events, the influence of the domain on the anticipation displayed in messages is an important factor that can not be ignored.

The goal of this research is to find out to what extent domains need to be considered when detecting prospective news events, and whether an overall anticipating pattern, independent of domain, is apparent. To this end, messages referring to scheduled or past events from different domains, posted through the social media platform of twitter.com, were collected and labeled as posted 'before', 'during' or 'after' the related event. In order to highlight similarities and differences between anticipating messages in different domains, a set of classification outcomes on different combinations of the data is presented. Included domains are football matches in the European Championships, the Lowlands music festival, the 2012 Dutch elections and a birthday party in Haren that spiralled out of control.

Semantic Frame Induction in an Assistive Vocal Interface using Hierarchical HMMs

Janneke van de Loo ✎
CLiPS, University of Antwerp, Antwerp, Belgium
Email: janneke.vandeloo@ua.ac.be

Jort F. Gemmeke
ESAT, KU Leuven, Leuven, Belgium
Email: jort.gemmeke@esat.kuleuven.be

Guy De Pauw
CLiPS, University of Antwerp, Antwerp, Belgium
Email: guy.depauw@ua.ac.be

Hugo Van hamme
ESAT, KU Leuven, Leuven, Belgium
Email: hugo.vanhamme@esat.kuleuven.be

Walter Daelemans
CLiPS, University of Antwerp, Antwerp, Belgium
Email: walter.daelemans@ua.ac.be

In the ALADIN project, we are developing a self-learning assistive vocal interface for people with physical impairments. This system should enable them to operate devices at home by giving vocal commands in an intuitive way; the system automatically learns the user-specific command structures and speech characteristics and adapts itself accordingly. The vocal interface is trained with a small number of training examples: spoken commands and their associated actions, of which the latter are represented as semantic frames with slots and fillers (slot values). The learning task of the system is to find meaningful units and structures in the spoken commands and relate them to slots and slot values in the semantic frames. After training, it should then be able to induce the semantic frame descriptions of commands uttered by the user.

We developed a semantic frame induction framework in which the spoken commands and their underlying semantic structures are modelled in hierarchical hidden Markov models (HHMMs). We present results of semantic frame induction experiments on command-and-control data which have been recorded in a voice-controlled card game setup. In these experiments, orthographic transcriptions of the spoken commands - both manual transcriptions and automatic transcriptions produced by a speech recognizer - are used as input for the semantic frame induction system. In future experiments, audio input will be used as well.

A "no-no" System (no Context, No Resource) for Classifying a List of Words into Correctly Spelled and Misspelled Items

Jean-Luc Manguin
CNRS & Université de Caen (France)
Email: jean-luc.manguin@unicaen.fr

In the last edition of CLIN, we described a "no-context" orthographic corrector which pairs the 15000 erroneous forms of a list of 58500 items with their correct forms, by use of different techniques : editing distance, phonetization, etc. But finding those misspelled items in the list was done with help of a lexical resource, so the purpose of the work described here is to classify the items of this list into two categories (correctly spelled vs misspelled) WITHOUT any resource. We only use statistical data concerning the most frequent relations in graphic neighbourhood, and the frequencies of the items. These data allow us to extract from the complete list a set of 24000 items in which these items are clustered in small groups. The elements of a group shares the same "graphic scheme" and are sorted by frequency. Then we assume that 1) the first element of each group is correctly spelled 2) the following elements are misspelled 3) the remaining 34500 items are correctly spelled. This method gives a 89 % precision and a 91 % recall (if we consider that the task is to extract correctly spelled items) and a 85 % global accuracy. Introducing a very small knowledge of the morphology can improve these results up to a 94 % recall (for correct items) and a 87 % accuracy. If we consider that the task is to extract the misspelled words, the precision then becomes 73 % and the recall 67 %.

Disambiguation of Word Translations without Parallel Corpora

Erwin Marsi ✎
Norwegian University of Science and Technology (NTNU)
Email: emarsi@idi.ntnu.no

Björn Gambäck
Norwegian University of Science and Technology (NTNU)
Email: gamback@idi.ntnu.no

Many words have more than one possible translation into another language. The proper translation depends on the context of use. Word translation disambiguation (WTD) concerns selecting the best translation(s) for a source word from a given set of translation candidates in a particular context. Machine translation systems are - implicitly or explicitly - faced with the task of WTD. Following the currently dominant paradigm of statistical MT, most systems address the task by learning from word-aligned parallel text. However, parallel corpora are a scarce resource and finding sufficient bilingual material is challenging for most many language pairs. Our work is therefore aimed at WTD models that do not rely on parallel corpora, but instead only require large monolingual corpora, which are more often readily available. The approach conditions models on the target language context and is completely unsupervised. One of the models explored is centroid-based classification, an instance of vector space models often employed in text categorisation. Experimental results are presented for several language pairs, indicating that the proposed WTD models result in improvement of overall translation quality. In addition, the potential, the challenges and the limitations of the approach are discussed.

Folktale Classification using Learning to Rank

Dong Nguyen ✎
University of Twente
Email: dong.p.ng@gmail.com

Dolf Trieschnigg
University of Twente
Email: d.trieschnigg@utwente.nl

Mariët Theune
University of Twente
Email: m.theune@utwente.nl

We present a learning to rank approach to classify folktales, such as fairy tales and urban legends, according to their story type, a concept that is widely used by folktale researchers to organize and classify folktales. A story type represents a collection of similar stories often with recurring plot and themes. Our work is guided by two frequently used story type classification schemes. Contrary to most information retrieval problems, the text similarity in this problem goes beyond topical similarity. We experiment with approaches inspired by distributed information retrieval and features that compare subject-verb-object triplets. Our system was found to be highly effective compared with a baseline system.

The Chatty Corpus: a Gold Standard for Dutch Chat Normalization

Claudia Peersman ✎
Antwerp University
Email: claudia.peersman@ua.ac.be

Mike Kestemont
Antwerp University
Email: mike.kestemont@ua.ac.be

Benny De Decker
Antwerp University
Email: benny.dedecker@ua.ac.be

Guy De Pauw
Antwerp University
Email: guy.depauw@ua.ac.be

Kim Luyckx
Antwerp University
Email: kim.luyckx@ua.ac.be

Roser Morante
Antwerp University
Email: roser.morante@ua.ac.be

Frederik Vaassen
Antwerp University
Email: frederik.vaassen@ua.ac.be

Janneke van de Loo
Antwerp University
Email: janneke.vandeloo@ua.ac.be

Walter Daelemans
Antwerp University
Email: walter.daelemans@ua.ac.be

In recent years, numerous forms of Internet communication, such as email, blogs, social network posts, tweets and chat room conversations, have emerged together with a new language variety called chat speak. In Flanders, chat speak does not only include Internet abbreviations (e.g. 'lol') and spelling errors, which are typical for chat speak, but also entails representations colloquial Flemish - a conglomerate of Dutch dialects spoken in the North of Belgium. This language variety differs significantly from standard (Netherlandic) Dutch, because it displays a lot more dialectal features, which are continued in chat speak. Because state-of-the-art NLP tools fail to correctly analyse the surface forms of (Flemish) chat language usage, we propose to normalize this 'anomalous' input into a format suitable for existing NLP solutions for standard Dutch. To achieve this, we have annotated a substantial part of a corpus of Flemish Dutch chat posts that were collected from the Belgian online social network Netlog in order to provide a gold standard for the evaluation of future approaches to automatic (Flemish) chat speak normalization. We discuss our annotation guidelines and inter-annotator agreement scores during our presentation. However, in a world where adolescent (chat) language varies constantly, we believe that machine learning approaches for the normalization of this text genre are needed which require minimal supervision, in order to reduce the cost and effort of manual annotation. Therefore, we also go into our normalization strategies we will investigate during our future research on this topic.

Using an Ontology to Reduce Verb Ambiguity

Marten Postma
Utrecht University
Email: M.C.Postma@students.uu.nl

he goal of this research is to observe which role the structured hierarchy of meanings inside the Dutch semantic ontology Cornetto (Vossen,2006) can play in reducing analysis ambiguity. The central issue is exemplified by (1) and (2):

(1) Hij nam Egypte in.
He took Egypte in
'He conquered Egypte.'

(2) Hij nam paracetamol in.
He took paracetamol in
'He swallowed paracetamol.'

Although the subject, tense, and syntactic construction in both examples are exactly the same, ambiguity doesn't arise in (1) and (2). However, the semantic classes of the direct objects in (1) and (2) are different, because Egypte is part of the semantic class of areas and paracetamol of the class of substances. Hence, we claim that if the direct object is part of the semantic class of areas, the meaning of the verb is to conquer and if the direct object is part of the semantic class of substances, the meaning is of the verb is to swallow. Areas and substances are so called semantic selection restrictions(SSR). We developed two WSD-systems that predict the meaning of a verb lemma based on the semantic class of the direct object. The first system trains itself on training data, from DutchSemCor (Vossen et al.,2011), to find a SSR per verb meaning and uses these SSRs to predict the meanings of verb lemmas in the test data. The second system is based on semantic similarity scores to perform cluster analysis with Cluto (Rasmussen & Karypis,2004). The results from both systems outperformed the baseline.

Building Realistic Spellnets for European Languages with Anagram Hashing

Martin Reynaert
TiCC - Tilburg University
Email: reynaert@uvt.nl

Experimental results in our previous work show that it is apparently harder to perform spelling correction on Dutch than on English. Choudhury et al. (2007) in "How Difficult is it to Develop a Perfect Spell-checker? A Cross-linguistic Analysis through Complex Network Approach" offer a tantalizing explanation that this is in fact due to the networking properties between words in a language. The network or Spellnet is then composed of the possible links between words definable as "word A is linked to word B if it is within edit distance n from B" where n is a small number of character edits. They construct SpellNets for three languages: Bengali, English and Hindi. The lexicons used consist of the alleged 10,000 most frequent words.

We replicate their work, aiming at the 11 European languages for which Google has provided the Web 1T 5-gram data. To be able to do this we employ our inexpensive solution for identifying these linked word forms: anagram hashing. To help scale this to realistic lexicon sizes we make use of the Dutch BigGrid. We first explain how we have calibrated our own implementation for measuring the relevant network properties, a.o. clustering coefficient, on the basis of the original English lexicon used by Choudhury. We contrast our findings given Spellnets built from all-wordform expanded open source spelling correction lexicons versus those built from the Google web frequency word lists. For Dutch, we also compare with the frequency list provided by the new contemporary, written Dutch reference corpus SoNaR.

Automatic Syllabification Using Segmental Conditional Random Fields

Kseniya Rogova ✎
KU Leuven, ESAT-PSI
Email: Kseniya.Rogova@esat.kuleuven.be

Kris Demuynck
Ghent University, ELIS
Email: Kris.Demuynck@UGent.be

Dirk Van Compernolle
KU Leuven, ESAT-PSI
Email: dirk.vancompernolle@esat.kuleuven.be

We present a statistical approach for the automatic syllabification of phonetic transcriptions of words. Fundamental to this work is the use of segmental conditional random fields (SCRFs). The SCRF models the conditional probability of a syllable sequence (labels) given the phonetic transcription of the word (the observations) as a log-linear combination of feature functions. A first set of features incorporates the main syllabification principles into the system by means of legality, sonority and maximal onset scores. The second set consists of syllable occurrence and co-occurrence statistics (bigrams) and consonant cluster split probabilities. Targeted applications of the approach include reading tutors, speech synthesis and speech recognition. Depending on the training data, we distinguish supervised, unsupervised and semi-supervised syllabifications. The approach is language-independent. The only language-dependent prior information is the set of phones and the sonority scale. Other information is derived from a corpus of data (with or without syllabification information). Phonetic syllabication rules and syllabification database show a fair amount of ambiguity. To address this problem, our model outputs several possible syllabifications with corresponding probabilities. The method was tested on two different datasets: CELEX and Wordsmyth, and on two languages: English and Dutch. The results presented are among the best published results so far, although comparisons are difficult due to the lack of standardization for the considered task. We obtained a 98.04% word accuracy for English supervised syllabification. Including the second best variant (needed for less than 15% of the words) increased the accuracy to 99.01%.

Dealing with Big Data: the Case of Twitter

Erik Tjong Kim Sang ✎
Netherlands eScience Center
Email: e.tjongkimsang@esciencecenter.nl

Antal van den Bosch
Radboud University Nijmegen
Email: a.vandenbosch@let.ru.nl

As data sets keep growing, computational linguists are experiencing more and more big data problems: challenging demands on storage and processing caused by very large data sets. An example of this is dealing with social media data: including the meta data, the messages of the social media site Twitter in 2012 comprise of more than 250 terabytes of structured text data. Handling data volumes like this requires parallel computing architectures with appropriate software tools.

In this talk we present our experiences in working with such a big data set, a collection of 1.5 billion Dutch tweets which make up 4 terabytes, including the meta data. We show how we collected and stored the data. Next we deal with searching in the data using the Hadoop framework and visualizing search results. We apply existing Dutch natural language tools to the data and test their performance on a parallel architecture.

Visualization of search results for such large data sets allows for interesting observations like daily patterns in medium usage and regional spread of dialect words. We will also present some of these observations.

Detecting Connected Text Using an Effective Unsupervised Method

Roland Schäfer
German Language and Linguistics, FU Berlin
Email: roland.schaefer@fu-berlin.de

I show that for a given language, a classification of documents as containing predominantly connected text can be achieved by examining relative frequencies of frequent words. It is based on an approach to language identification first described by Grefenstette in 1995 and the method used by the WaCky team. Supervised methods bring about the problem of the virtually impossible operationalization of "well-connected documents". Especially web corpora contain mixtures of good text, tag clouds, foreign material, etc., requiring arbitrary individual decisions between good and bad documents. Therefore, my approach simply measures the relative frequencies for the n most frequent types in a training set. Frequent words are usually function words, which are characteristic of connected text. The resulting classifier removes documents where the frequencies of these types deviate too strongly (to the negative side) from the measured (weighted) means. The cutoff is simply specified as a percentile of the measured distribution. The classifier thus just homogenizes the language in the corpus with adjustable strength, including the removal of non-connected text. In experiments (Dutch, English, French, German, Malay, Swedish), corpora generated from search engine-returned URLs queried with 4-tuples of mid-frequency words provide a stable training input above a size of 100-200 documents. To avoid subjective decisions, the method is evaluated on artificial documents mixed from doubtlessly connected text and tag cloud material. It is shown to be reliable above a document size of roughly 200 words. As a by-product, it works as a simple language identifier with Precision and Recall above 0.99.

Simulating the Acquisition of Dutch Word-initial Consonant Clusters with Q-HG

Klaas Seinhorst ✎
University of Amsterdam
Email: seinhorst@uva.nl

Tamas Biro
University of Amsterdam
Email: t.s.biro@uva.nl

Early stages of linguistic development are characterized by an array of simplification phenomena. In phonology, Dutch children often omit one or more segments in word-initial consonant clusters. For instance, they may produce the word *bloem* 'flower' as *boem*, or *slapen* 'to sleep' as *lapen*, before mastering the target form. Within an Optimality-Theoretic framework this phenomenon is usually explained as the result of constraint reranking: the child gradually learns that faithful production of the target form is more important than the markedness of phonological complexity.

We take an alternative stance and posit instead that the child acquires the correct knowledge without delay, but is not yet able to compute the target form correctly. We present the preliminary results of computer simulations of child language corpus data. Our simulations aim at reproducing the distinct learning curves of four different types of word-initial consonant clusters. We approximate OT with Harmonic Grammar with weights that are powers of a base q (a q-HG grammar), and postulate that q increases as a function of age. This way, the grammar comes to behave as an OT grammar by a certain age: with low values of q, the grammar produces the simplified form, whereas with high values of q, it produces the adult form.

In order to better reproduce the different learning curves, we also experimented with probabilistic variants of OT, such as Boersma's Stochastic OT and Biró's Simulated Annealing. We found the most promising results with Simulated Annealing and a restricted topology.

Novel Developments in ELAN

Han Sloetjes
Max Planck for Psycholinguistics
Email: han.sloetjes@mpi.nl

Herman Stehouwer ✎
Max Planck for Psycholinguistics
Email: herman.stehouwer@mpi.nl

Sebastian Drude
Max Planck for Psycholinguistics
Email: sebastian.drude@mpi.nl

Modern language documentation depends on suitable software infrastructure. ELAN is a well-known tool developed at The Language Archive / MPI-PL which allows multi-tier, multi-speaker, time-linked annotation of audio and video recordings, in particular in a field work and language documentation setting.

In the past two years ELAN has been under constant development. Here we will give an overview of the major recent enhancements to ELAN and ongoing work. Below we address five aspects, each consisting of multiple new features.

First, there are modes that help you perform specific tasks. These are the segmentation, and the transcription mode. With a user interface for each task, the modes together provide very efficient means for the initial steps of a typical workflow.

Second, the interoperability with FLEx (FieldWork Language Explorer) has been improved.

Third, extensive support multiple-file operations have been added. These include a) file-format conversion, and b) creation of similarly structured EAF files for media files.

Fourth, facilities have been added to create new tiers with annotations on the basis of existing tiers with logical operations. E.g. if the annotation occurs in both tier A and tier B, then copy it combined to tier C. These features provide for a straightforward interface to basic, but extremely helpful operations.

Fifth, preliminary interaction with relevant web services (online audio-video and text processors that create annotations) has been implemented.

In short, in the past years several crucial features have been added that make ELAN better and faster to use in many aspects.

Under the Hood of an English Context-sensitive Spell Checker

Wessel Stoop ✎
Radboud University Nijmegen
Email: wesselstoop@student.ru.nl

Antal van den Bosch
Radboud University Nijmegen
Email: a.vandenbosch@let.ru.nl

Maarten van Gompel
Radboud University Nijmegen
Email: proycon@anaproy.nl

Peter Berck
Tilburg University
Email: pberck@uvt.nl

We explain the inner workings of fowlt.net, an online spell checker freely available on the internet. Fowlt.net, just like it's Dutch equivalent valkuil.net, is a context-sensitive spell checker, using the memory based learning classifier Timbl and word prediction software WOPR. The general idea is that a certain word is tagged as an error if our language models expected another word there and are very certain about it. This makes detection of errors like the one in the second sentence of this abstract possible. Fowlt consists of multiple modules, all approaching the task of detecting and correcting errors from different viewpoints. We will explain how we train and tune these modules, and how they operate together. We provide preliminary and small-scale evaluations and comparisons to rival free online systems for English spelling and grammar checking.

Linear Model for Exploring Types of Vowel Harmony

Lili Szabó ✎
Universität des Saarlandes
Email: lilis@coli.uni-saarland.de

Çağrı Çöltekin
Rijksuniversiteit Groningen
Email: c.coltekin@rug.nl

Vowel harmony (VH) is a phonological phenomenon observed in some languages where vowels agree/harmonize with other vowels within a given scope, according to one or more phonological features such as backness, highness, roundedness. Once speakers of these languages discover the type of VH their languages exhibit (e.g. backness-harmony), they use this knowledge to select the correct allophones in suffixation, or to segment words in continuous speech.

This study describes a statistical model that discovers whether a language exhibits VH, and if it does, which type of it. Our model exploits the relationship between co-occurrence of non-adjacent consecutive vowels and their acoustic properties. We use a linear model that predicts the co-occurrence of two vowels from their acoustic properties, where the co-occurence is measured by pointwise mutual information (PMI). Both PMI and the acoustic properties are automatically extracted from unannotated corpora.

We tested our model on multiple languages which exhibit (Hungarian, Turkish) and which don't exhibit VH (English, Dutch). Our results confirm the earlier findings in the literature. In addition, we also show for languages which exhibit VH that (1) PMI-scores are significantly higher for harmonizing vowel pairs; (2) there is interaction among two harmony types (backness and roundedness); (3) the model is able to explain around half of the variance for languages with VH.

Although the model is intended as an interactive analysis tool, the results can also be useful in explaining how children learn VH, it can guide to better models in speech processing and unsupervised morphological analysis.

A Simple Method for Topic Classification for Morphologically Complex Languages

Frederik Vaassen ✎
CLiPS Research Center, University of Antwerp, Belgium
Email: frederik.vaassen@ua.ac.be

Jurgita Kapočiuė-Dzikienė
Kaunas University of Technology, Lithuania
Email: Jurgita.Kapociute-Dzikiene@ktu.lt

Guy De Pauw
CLiPS Research Center, University of Antwerp, Belgium
Email: guy.depauw@ua.ac.be

Walter Daelemans
CLiPS Research Center, University of Antwerp, Belgium
Email: walter.daelemans@ua.ac.be

While using a simple bag-of-words approach is a common and efficient method for topic classification of English text documents, it is not guaranteed that this method is also appropriate for languages that are substantially different. Through solving a topic classification task for Lithuanian -a relatively resource-scarce language that is highly inflective, has a rich vocabulary, and a complex word derivation system- we show that it is possible to significantly improve on the common bag-of-words approach by taking the morphological complexity of the language into account. A classifier based on character n-grams will be able to focus on those character strings that are semantically salient for a given topic, thus making abstraction of inflectional or derivational affixes which would only confuse a bag-of-words classifier. We show that these findings hold for three different datasets, with different topic distributions and containing both formal and informal Lithuanian text. We demonstrate that a simple character n-gram approach even surpasses bag-of-words classifiers based on stemmed or lemmatized text. This conclusion is especially important since it shows that topic classification is feasible for resource-scarce languages where stemmers or lemmatizers are not readily available. Finally, we hypothesize that the character n-gram method can be applied to any language with a similarly complex morphology. To confirm this hypothesis, we carry out a series of experiments on two additional languages: Russian and Swahili. While otherwise unrelated, Russian and Swahili are both highly subject to affixation and complex word derivations, which makes them ideal targets for the character n-gram approach.

Using Natural Language as a Bridge Between Pictograph Sets

Vincent Vandeghinste
Centrum voor Computerlinguïstiek - KU Leuven
Email: vincent@ccl.kuleuven.be

The WAI-NOT environment is a platform which allows people with cognitive disabilities to communicate online using pictographs instead of text and supports two pictograph sets. Users can enter messages using their pictograph set and/or text. These messages are encoded as text and sent to the receiver where they are decoded into the target pictograph set wherever possible.

There are two problems with this approach: 1. The decoding is purely string-based and no disambiguation takes place, which occasionally leads to wrong pictograph generation; and 2. The current string-based overlap between the two sets is too small to be of practical usage

We have collected a corpus of 200K words of e-mail messages sent with WAI-NOT, and we show how simple NLP techniques such as part-of-speech tagging and lemmatisation can improve the conversion of messages from one pictograph set to the other. A relative improvement of $> 45\%$ was reached on unseen data. Furthermore we discuss how in the next phase we will use word-sense-disambiguation and linking to Cornetto, a lexical-semantic database, to further improve the results.

Automatic Thematical Classification of Party Programmes

Suzan Verberne ✎
Radboud University Nijmegen
Email: s.verberne@let.ru.nl

Eva D'hondt
Radboud University Nijmegen
Email: e.dhondt@let.ru.nl

Antal van den Bosch
Radboud University Nijmegen
Email: a.vandenbosch@let.ru.nl

Maarten Marx
University of Amsterdam
Email: maartenmarx@uva.nl

Isaac Lipschits was a Dutch historian and political scientist. One of his works is the annotated collection of party programmes for the Dutch elections (1977-1998). For each election year, he compiled a book with the programmes for all parties, he segmented the documents in coherent text fragments and labelled them with themes.

In the PoliticalMashup project, Dutch political data from 1946 onwards is being digitized, among which are the annotated party programmes by Isaac Lipschits. We aim to (1) digitize the 1977-1998 Lipschits collections and (2) build an automatic classifier for more recent, unclassified editions.

In our first task, the digitization of the party programmes, we encountered a number of complicating OCR errors, which we will discuss in our presentation. The second task, i.e. training a classifier on the annotated data, was challenging because the texts are short (about 300 words on average) and Lipschits assigned more than 6 themes per text on average, and more than 200 different themes in total. To build the automatic classifier, we trained several multi-label classification models on the labelled data from the 80's and 90's, optimizing on the most recent data (1998). We achieved a precision of 75% and a recall of 40% - relatively low due to the large differences in themes between the individual years.

After optimization, we used our classifier to assign themes to the programmes from 2006 onwards. A sample of the labels was evaluated by an expert. We will show and discuss the results.

Semantic Classification of Dutch and Afrikaans Noun-Noun Compounds

Ben Verhoeven ✎
CLiPS, University of Antwerp
Email: ben.verhoeven@student.ua.ac.be

Walter Daelemans
CLiPS, University of Antwerp
Email: walter.daelemans@ua.ac.be

Gerhard van Huyssteen
CTexT, North-West University
Email: gerhard.vanhuyssteen@nwu.ac.za

The meaning of compound words is often ambiguous because there is no explicit description of the relation between the compound constituents. A newly produced compound like 'donut seat' can be interpreted in different ways, such as 'seat with donut nearby', 'seat that looks like donut' or even 'seat made of donuts'. An automatic semantic analysis of these compounds may shed more light on this issue.

Building on previous research by Ó Séaghda for English, the task of semantically analysing noun-noun compounds was considered a supervised machine learning problem. We adopted and adapted Ó Séaghda's semantic classification scheme and guidelines for noun-noun compounds. This scheme describes 11 classes, of which 6 are semantically specific. Lists of noun-noun compounds were annotated according to this classification scheme. Following the distributional hypothesis that states that the set of contexts of a word can be used as a representation of its meaning, vectors with co-occurrence information on the compound constituent nouns were used to construct feature vectors for our classifier. We present results of our experiments on Dutch and Afrikaans compounds that confirm the learnability of this classification task. Different experiments vary in the number and kind of co-occurrence words they select (e.g. content words vs. function words).

Our results are promising and approach the accuracies reached by similar systems for English.

A Graph-based Approach for Implicit Discourse Relations

Yannick Versley
Univ. of Tübingen
Email: versley@sfs.uni-tuebingen.de

Recognizing implicit discourse relations, such as causal relations (Explanation, Result) and those signaling thematic expansion (Background, Instance, Restatement), is an important ingredient in the structural analysis of text beyond single clauses or sentences. Current machine learning approaches for this task (Sporleder and Lascarides, 2008; Lin et al., 2009) heavily rely on shallow features such as word pairs or bigrams, and attempts to include linguistic features have met limited success. We use graph-based representations of clauses as a means to integrate linguistic and structural information in a manner suitable for modern machine learning algorithms. Using a corpus containing (among other) 803 implicit discourse relations in German newspaper articles from the TüBa-D/Z corpus (Gastel et al., 2011; Telljohann et al., 2012) we show that a graph-based approach of integrating linguistic, lexical, and structural information performs well against several strong shallow-feature baselines involving bigrams, word pairs, or grammar productions, and also outperforms a system only using linguistic features.

References:

Gastel, A., S. Schulze, Y. Versley & E. Hinrichs (2011). Annotation of implicit discourse relations in the TüBa-D/Z treebank. In Proc. GSCL 2011.

Lin, Z., M. Y. Kan & H. T. Ng (2009). Recognizing implicit discourse relations in the Penn Discourse Treebank. In Proc. EMNLP 2009.

Sporleder, C. & A. Lascarides (2008). Using automatically labeled examples to classify rhetorical relations: An assessment. Natural Language Engineering 14:369–416.

Telljohann, H., E. Hinrichs, S. Kübler, H. Zinsmeister, K. Beck (2012). Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical Report, Sem. f. Sprachwissenschaft, Univ. Tübingen.

Unsupervised Pattern Discovery in Speech

Maarten Versteegh ✎
Radboud University / International Max Planck Research School for Language Sciences
Email: m.versteegh@let.ru.nl

Michele Gubian
Radboud University
Email: m.gubian@let.ru.nl

Lou Boves
Radboud University
Email: l.boves@ru.nl

Recent investigations in computational modeling of first language acquisition and applications of speech technology to under-resourced languages have pointed out the similarity between these areas. Both aim their attention at the same class of acoustic pattern matching algorithms, which in the former case are used to simulate the discovery of word-like units in a child's mind, while in the latter case are employed to isolate potentially meaningful acoustic units from large untranscribed databases.

The work we present here is a novel approach to speech processing that attempts to discover recurrent patterns in a speech stream in an unsupervised way. We employ a modification of the well known Dynamic Time Warping algorithm and a data driven distance-to-similarity transformation to find recurrent stretches in speech streams. Building on these two advances, we show that our algorithm can reliably detect word-like units in speech utterances in an efficient manner.

The ability to discover units of speech in an unsupervised manner can find applications in word segmentation, keyword spotting and discovery, and lexical inventory building for under-resourced languages as well as help in our understanding of how infants learn the words and sounds of their language.

Referring Expression Generation in 3D Scenes: Modelling Human Reference Production using Graphs

Jette Viethen ✎
Tilburg University
Email: h.a.e.viethen@tilburguniversity.edu

Emiel Krahmer
Tilburg University
Email: e.j.krahmer@tilburguniversity.edu

Referring expression generation (REG), a task which is usually understood to consist of choosing the attributes of a target referent that should be used in a full referring noun phrase, forms an important part of any Natural Language Generation system, as well-chosen referring expressions are crucial to ensuring successful communication. Existing REG algorithms have typically been conceived for and evaluated on very basic scenarios. In particular, the the data used in the recent REG Challenges (Gatt et al. 2009), which assessed system output for human-likeness, did not contain any referring expressions that make use of relations between the target referent and other entities. However, the Graph-based Algorithm by Krahmer et al. (2003), which was one of the best performing algorithms in these Challenges, uses a data representation which seems particularly well-suited for dealing with relations between objects. In the present work, we therefore make use of the GRE3D3 Corpus (Viethen and Dale 2008) in order to test the Graph-based Algorithm's in-principle capability of generating the kinds of relational descriptions that people use in simple 3D domains. This investigation shows that as soon as relations between objects are involved, the number of informationally redundant attributes that the algorithm can include in a referring expression is limited. More specifically, the algorithm is not able to include as much information about the landmark object as humans usually do. Following this, we discuss the advantages and disadvantages of a number of different solutions to this problem.

Toward a Model for Incremental Grounding in Dialogue Systems

Thomas Visser
University of Twente
Email: thomas.visser@gmail.com

Spoken dialogue systems have until recently upheld the simplifying assumption that the conversation between the user and the system occurs in a strict turn-by-turn fashion. In order to have more human-like, fluent conversations with computers, a new generation of spoken dialogue systems has arisen that is capable of processing the user's speech in an incremental way.

We have studied the AMI Meeting Corpus in order to identify ways of grounding in human-human dialogue that a system would be able to pick up using incremental processing. These incremental grounding behaviors include overlapping feedback, the completion of unfinished utterances that were initiated by the other party and responding to an utterance before it is completed.

We have developed an incremental grounding model that supports those incremental grounding behaviors. The input of the model consists of incremental hypotheses of the explicit and predicted content of the utterance, i.e. what has been uttered so far and what is likely to be the full utterance meaning respectively. We have defined how grounding acts can be identified incrementally and how the grounding state, i.e. the collected contents and progress of the common ground units (CGUs), is updated accordingly. We defined new types of acknowledgments and how they affect the content of the CGU they ground, e.g. answering an unfinished question also grounds the part of the question that was not uttered. We implemented our model in the SASO4 dialogue system as a proof-of-concept of our approach, showcasing an up-to-date grounding state through the execution of a simple overlapping feedback policy.

DutchSemCor: in Quest of the Ideal Dutch Semantic Corpus

Piek Vossen
VU University Amsterdam
Email: p.vossen@vu.nl

Rubén Izquierdo
VU University Amsterdam
Email: izquierdo@let.vu.nl

Attila Görög ✎
VU University Amsterdam
Email: a.gorog@vu.nl

Sense-tagged corpora should preferably represent all senses of words including rare senses, the variety of contexts and provide information on sense distribution. These three requirements are often contradictory in practice. In this presentation, we will describe our approach of trying to meet these three requirements in the NWO project DutchSemCor.

The goal of DutchSemCor was to deliver a Dutch corpus that is sense-tagged with senses and domain tags of the 3.000 most frequent and polysemous Dutch words from the Cornetto database. More than 300K words were manually tagged by at least two annotators. The data was used to train and evaluate 3 WSD-systems: TiMBL-DSC, SVM-DSC and UKB-DSC. UKB-DSC was trained using circa 1.8 million relations (partly derived from Cornetto and WordNet and partly from the manually-tagged data). The systems were tested in three independent evaluations: fold-cross, all-word and random evaluation (with the best result of 88,65 reached in Fold-cross evaluation).

At the end of the project, all tokens for the selected lemmas in the Dutch SoNaR corpus were automatically sense-tagged using a combination of these WSD-systems and assigning a confidence value to each token. A portion of the automatically annotated data was validated by human annotators.

The corpus data is based on existing corpus material and was extended using web-snippets to find sufficient examples for less frequent senses. The resulting corpus is extremely rich in terms of lexical semantic information. Its availability enables many new lines of research and technology developments for the Dutch language.

Interpreting Intentions of Speech

Ke Wang ✎
Dalian University of Technology
Email: wang.coco.ke@gmail.com

Gerald Penn
University of Toronto
Email: gpenn@cs.toronto.edu

The study of intentions of speech arises from the researches in linguistics, philosophy, and more recently, artificial intelligence, robotics, and psychology. Since the 80s and 90s of last century, many scientists (P. Cohen et al. 1990) have tried to interpret intentions through logic inference. However, lacking of effective means of semantic analysis, the results are not very satisfying. Vanderveken (1990) tried to construct a logic of illocutionary force. But he couldn't reveal the semantic implications, because he couldn't explain how different illocutionary forces compose as a whole. The purpose of this paper is to reveal the mechanism of interpreting intentions of speech. Firstly, we present a grammar system for extracting the semantic structures of intentions. The grammar system includes a taxonomy of intentions of speech which is based on Speech Act Theory and Searle's philosophy about "social reality", and a set of grammar rules. Then, we give a logic of semantic implication that explains how people understand and respond to the implicit meanings of complex intentions, such as an imperative hidden in a query, and a query embedded in indirect speech in (1) and (2) respectively. (1) Could you bring me the book? (2) He asked if you would come.

Metadata for Tools: a CMDI Software Profile for Describing Your Software

Eline Westerhout ✎
Universiteit Utrecht
Email: e.n.westerhout@uu.nl

Jan Odijk
Universiteit Utrecht
Email: j.odijk@uu.nl

In the last couple of years, the CLARIN-NL project has supported several resource curation and demonstration projects related to humanities topics and disciplines in The Netherlands. As a result, a wide range of tools and resources have become available for researchers and users. However, in order to maximize the reusability potential and the consistency of the software descriptions, the creation of standardized metadata profiles is of crucial importance.

In this talk, we will present the CMDI metadata profiles for describing software. Depending on the target group, two types of profiles have been created. The first profile is intended for users of the software and provides a general description of the software function, usage and availability (user-oriented). The second profile aims to capture development and documentation information and describes the technical characteristics of the software (developer-oriented). A link will be established with the Virtual Language Observatory (VLO). All elements and components have been linked to ISOcata Data Categories.

We will demonstrate what input and efforts are required from the CLARIN-NL projects to create metadata for their software according to the user-oriented and developer-oriented profiles. Other software, that is not part of the CLARIN-NL architecture yet, can be described using these profiles as well.

Industrial Track

CLIN 2013

Why Multi-lingual Associative Networks Are Better At Categorization Than Their Monolingual Equivalents

Niels Bloom
University of Twente, Pagelink Interactives
Email: n.bloom@perrit.nl

Associative networks can use raw text to categorize documents intuitively and accurately. To do this, they require basic linguistic information about the relationships between words and concepts, which can be extracted from, for example, Wikipedia or Princeton WordNet.

However, associative networks are not limited to a single language - instead, they can combine linguistic information from multiple languages, allowing them to categorize documents in different languages or even documents which contain texts in multiple languages. Moreover associative networks can categorize such documents correctly without a need to translate them, as terms in one language are linked to similar terms in a different language within the associative network, allowing it to make connections between the documents across the language barrier and in turn using this to discover which articles should be grouped together.

Because of those same cross-lingual links, multilingual associative networks are able to get better results than mono lingual ones, using the new connections through the different language to focus activation of the network into relevant clusters, as well as recognizing foreign words within a text with ease.

Matching CVs to Job Vacancies: Semantic Search in Action

Remko Bonnema
Textkernel
Email: bonnema@textkernel.nl

Gerard Goossen
Textkernel
Email: goossen@textkernel.nl

Mihai Rotaru
Textkernel
Email: rotaru@textkernel.nl

Lena Bayeva
Textkernel
Email: bayeva@textkernel.nl

Chao Li
Textkernel
Email: chaoli@textkernel.nl

Florence Berbain
Textkernel
Email: berbain@textkernel.nl

Carsten Lygteskov Hansen ✎
Textkernel
Email: hansen@textkernel.nl

Document understanding systems have proven very successful within many different domains by transforming unstructured documents into a structured and searchable information source. However, a major problem still remains for users: finding the desired information requires both knowledge of the domain and of the query language. A successful solution should guide and inspire the user' search without enforcing unwanted assumptions. One example of such assumption is to expect users have a full grasp of the data model used for structuring the CVs when searching for suitable candidates in a database of analyzed CVs.

Our solution to the above problem is to automatically transform a job vacancy into a query. This query is then used to search for matching CVs and it is fully compliant with the data model used in the CVs database. Such queries allow for optimal results even for inexperienced users. The whole process employs a wide variety of NLP techniques which are required to bridge the lexical and semantic gaps between the job vacancy and the CV domains.

The talk consists of a demonstration of the system and a presentation of the different features available. We also present an overview of the steps taken to improve the matching results and a discussion of what is still needed to find an optimal match between job vacancies and CVs.

Charting the Noosphere: Mapping Internal Realities to External Worlds

ML Coler
INCAS^3, Assen
Email: mattcoler@incas3.eu

This project establishes the relation between the multifaceted arrays of human sensory experience and knowledge with physical phenomena to understand how people effect, interact with, and create environments, even predicting how an area is used. In collaboration with cognitive scientists, information scientists and representatives from industry and the municipality of Assen, we will chart something like Vernadsky's noosphere; i.e., the layer of thought and reason enveloping the world (in more modern terms: the intertwined flux of physical and mental phenomena that shape behavior).

Particularly important is the role of cognitive tests and language. All spoken language data, once transcribed, will be automatically annotated and subjected to syntactic evaluation with a dependency parser. Thereafter, frames of reference relevant to specific aspects of human experience are identified. To the extent possible, these frames are mapped to the output of the dependency parser as patterns emerge. A subset of the data then undergoes cognitive semantic analysis, aided by psycholinguistic tests to uncover tacit category structure.

This research will add a layer to our maps, going beyond the streets, houses, and rivers to provide the essence of human experience: activity and behavior. Understanding how people use an area based on a physical analysis of the acoustic and visual environment coupled with ethnographic and linguistic techniques to get at the nature of experience and expectation will drive the development of intelligent sensors capable of linking the quantifiable external world and the internal complexities of the mind to maximize our environments.

How to Search Annotated Text by Strategy?

Roberto Cornacchia
Spinque, Utrecht, The Netherlands
Email: roberto@spinque.com

Wouter Alink
Spinque, Utrecht, The Netherlands
Email: wouter@spinque.com

Arjen P. de Vries ✎
Spinque, Utrecht, The Netherlands
Email: arjen@acm.org

Search by strategy refers to an iterative 2-stage search process that separates search strategy definition (the how) from the actual searching and browsing of the collection (the what). The extra flexibility gained by this separation allows to vary an information retrieval system's internal ranking model dependent on the search problem at hand. This is especially useful for information access problems where we need to go beyond classic document retrieval needs; this talk will focus on the case where the document representation is annotated text.

We discuss how annotated text is represented internally, and how search strategies expressed graphically are transformed into efficient database query plans. We conclude with a preliminary example case from a children's information need, to illustrate how search strategies that deploy NLP features may improve over the traditional plain text retrieval approach.

Virtual Assistants: New Challenges to Language Technology

Christian Kissig ✎
Oracle Nederland BV
Email: christian.kissig@oracle.com

Fabrice Nauze
Oracle Nederland BV
Email: fabrice.nauze@oracle.com

Mandy Schiffrin
Oracle Nederland BV
Email: mandy.schiffrin@oracle.com

Virtual assistants (VAs) are computer generated agents used on websites, mobile applications, or kiosks which provide conversational interaction to respond to users' informational needs, guide users through processes, or perform requested actions. Since the introduction of Siri, S Voice, as well as various Android based implementations of intelligent personal assistant software into the mainstream consumer market, have VAs gained unprecedented attention in the information industry.

Natural language processing (NLP) figures prominently in generating the human-like experience the VA provides. NLP is used to process user input as well as to formulate responses. Further, as VAs operate on increasingly sensitive information, for instance in processing financial transactions, NLP algorithms have to satisfy high standards of quality.

In our work we have realized that although NLP is well understood in a static context, little is known about development metrics and evaluation in the context of dialogues. We describe particularities of NLP in conversational scenarios with history, memory and conversational context, outline some challenges for NLP for real-world VA implementations, and propose a roadmap to address some of these challenges.

Controlled Automatic Translation with Euroglot™

Leo Konst
Linguistic Systems BV
Email: leokonst@euroglot.nl

It will be argued that full automatic translation still is far from being possible and may never be possible. In the meantime we can build MT-programs within a controlled environment, where the user can verify the quality of the translation and can correct or change the translation easily. Euroglot™is such a program and we will discuss and demonstrate it. The program is multilingual, rule-based and uses the intelligent and intuïtive dictionnaries of Euroglot™Professional.

ZieOok: A Recommendation Platform for the Netherlands Institute for Sound and Vision

Oele Koornwinder ✎
GridLine
Email: oele@gridline.nl

Job Tiel Groenestege
GridLine
Email: job@gridline.nl

Many cultural heritage institutes have large collections that they want to share with the public. But when exploring such collections, visitors often get overwhelmed by the amount of data, making it hard for them to find items that are of interest to them.

Recommender systems can help people find content of their interest, showing items that are similar to items they viewed, or were viewed by people with a similar profile. Commissioned by the Netherlands Institute for Sound and Vision, GridLine developed ZieOok: a platform for creating, training, managing and consulting recommenders. Content providers can link their collection to the system through standard protocols, create and train a recommender for their collection and embed a recommendation widget in their collection's website.

ZieOok produces user-based and item-based recommenders. The first type generates recommendations for a single user, based on that user's profile through collaborative filtering. Item-based recommendations are based on similarity between items. GridLine incorporated various NLP techniques for improving item-based recommendation.

ZieOok is open source and is built on top of the open source machine learning framework Apache Mahout.

Automated Readability Scoring for Dutch Texts with SVMs

Roelant Ossewaarde
Edia b.v.
Email: roelant@edia.nl

For reading instruction, there is a need to determine whether a given text is appropriate for a given student level ("text readability"). Textual features ranging from the morphological to the discourse domain have been proposed as good predictors for readability.

Existing approaches use a single set of weighted features to predict text readability (Staphorsius (1994), Staphorsius & Krom (2008), Kraf & Pander Maat (2009)), in the form of a formula that yields a score ("single formula approach"). Most recently, Kraf & Pander Maat (2009) found only a handful of textual features with significant correlation to text readability.

In the single formula approach, the relation (weights) between regressors is constant across the range of the linear function. We challenge this generalization: discriminating between levels easiest and easy may require different parametrization than discriminating between levels hard and hardest.

We present an approach based on Support Vector Machines (SVM, Vapnik & Cortes (1995)) with features selected through heuristic hill climbing and backward elimination. Similar approaches have been successful for English, Swedish, Arabic, Japanese inter alia.

Compared to the single formula approach, our approach performs better (higher Jaccard coefficient) and identifies more features that correlate significantly with text readability. Some of these are intra-textual ones, whose correlation depends on their position in the text.

Posters & Demonstrations

CLIN 2013

Lexical Association Analysis For Semantic-Class Feature Enhancement In Parsing

Simon Šuster ✎
Rijksuniversiteit Groningen
Email: s.suster@rug.nl

Gertjan van Noord
Rijksuniversiteit Groningen
Email: G.J.M.van.Noord@rug.nl

We present a method for analysing the lexical association component of the disambiguation part of the Alpino parser for Dutch. The current component uses numerous features modelled from a 500M word corpus. For some dependency instances, the model offers no information, thus not contributing to a potentially correct parse analysis. An approach enhancing the features with more abstract semantic classes could overcome the drawbacks of lexical-association modelling based on lemma counts, and possibly lead to a further parsing improvement.

We thus analyse the performance of the parser for a particular feature in order to identify constructions where the lexical-association model could be improved. We note that there is indeed room for improvement of the parser both in the lexical-association aspect and in general parsing accuracy; that the exclusion of lexical-association information negatively correlates with the parser performance; that some features provide information only for a fraction of dependency instances. Based on this, we establish a list of the most promising features, i.e. the ones that are most likely to improve parsing after being enhanced with semantic classes in the future.

Faster Text Search with Hybrid Indexing

Eric Auer
The Language Archive, Max Planck Institute for Psycholinguistics, Nijmegen
Email: eric.auer@mpi.nl

Growing amounts of annotation data in The Language Archive make it necessary to significantly speed up search to keep response times user friendly. Unlike keyword oriented web search engines, the Trova and CQL Search services at TLA allow searching for arbitrary exact substrings and (at lower speed) even regular expressions, not just whole words.

To achieve both fast and versatile search, a combination of indexes is used. Word, substring and regular expression search queries are analyzed, yielding information about substrings and other properties which must be present in a tier (or file) so that tier can contain a hit for the query in question at all.

Those properties are then either hash-mapped to fixed size bit vectors (fingerprints) for PostgreSQL based filtering or expressed as sets of N-grams (up to a fixed length) for filtering with Lucene N-gram indexes.

Both methods aim to quickly find a small list of candidate tiers, containing all (but not much more) tiers which may contain hits. As Lucene has no native support for substring search, our system uses a fast but accurate N-gram based approximation.

We present details of the implemented algorithm and elaborate the improvements in response times achieved. We were able to speed up most steps (of: opening indexes, defining a search domain, gathering candidates, finding hits and collecting hit details) and a typical benchmark session now completes in a fraction of the time used by the already powerful previous implementation.

"Dit Het Op(ge)hou Reën." The IPP Effect in Afrikaans.

Liesbeth Augustinus ✎
CCL, KU Leuven
Email: liesbeth@ccl.kuleuven.be

Peter Dirix
CCL, KU Leuven
Email: peter@ccl.kuleuven.be

'Infinitivus pro participio' (IPP) is a syntactic phenomenon occurring in a subset of West Germanic languages, such as Dutch, German, and Afrikaans. It refers to constructions with a perfect auxiliary, in which an infinitive appears instead of a past participle. An Afrikaans example is 'Dit het ophou reën' (Eng. 'It has stopped raining').

The phenomenon has been studied extensively for Dutch and German, but studies on Afrikaans IPP triggers are sparse. Besides the IPP effect as it appears in Dutch and German, Donaldson (1993) mentions in his grammar of Afrikaans a second IPP construction: the perfect of progressive constructions. In those constructions the first verb may be represented as an infinitive or a past participle, e.g. 'Ek het soms buite (ge)sit en luister' . (Eng. 'Sometimes I sat outside and listened').

In order to obtain a clearer picture of both phenomena, we automatically created a treebank version of the Taalkommissie corpus. In order to do this, we developed a tokenizer and an Afrikaans version of the ShaRPa parser. We used the CTexT POS tagger for Afrikaans.

In contrast to Dutch and German, which have several verbs that obligatorily trigger IPP, it is often assumed in the literature that in Afrikaans IPP occurs optionally. The corpus results show, however, that some verbs have a clear preference for IPP, whereas other verbs prefer the corresponding construction with a past participle. Finally, a cross-linguistic comparison with Dutch provides support for making generalisations on verbs occurring in IPP constructions.

Wordrobe: using Games with a Purpose for Linguistic Annotation

Valerio Basile
University of Groningen
Email: v.basile@rug.nl

Johan Bos
University of Groningen
Email: johan.bos@rug.nl

Kilian Evang
University of Groningen
Email: k.evang@rug.nl

Noortje Venhuizen ✎
University of Groningen
Email: n.j.venhuizen@rug.nl

One of the main problems for statistical approaches in natural language processing is the need for extremely large amounts of data. Collecting this data is expensive, as it is a time-consuming task often performed by paid annotators. Moreover, there is a high need for quality control, since corrupted data may lead to low performance of linguistic tools. In the context of constructing a large corpus of semantically annotated texts, the Groningen Meaning Bank (Basile et al., 2012), we propose to tackle this problem by making use of crowdsourcing via a 'Game With A Purpose', called Wordrobe (www.wordrobe.org). GWAPs challenge players to score high on specifically designed tasks, thereby contributing their knowledge in an entertaining way. Wordrobe is a collection of linguistic games, focusing on different levels of linguistic analysis, including part-of-speech tagging, word-sense disambiguation, named entity recognition and coreference resolution. GWAPs have already been successfully employed in NLP-related research with initiatives like 'Phrase Detectives' (Chamberlain et al., 2008) and 'Jeux de Mots' (Artignan et al., 2009). While these games focus on specific phenomena, our aim is to cover different levels of linguistic analysis using a collection of games that share the same structure. The collected annotations ultimately enrich and improve the Groningen Meaning Bank, which is also used to automatically generate the game material. We will present the design and use of Wordrobe, as well as the first annotation results.

Syntactic Analysis of Arabic Coordination with HPSG Grammar

Sirine Boukédi ✎
Faculty of Science Economy and management
Email: sirine.boukedi@gmail.com

Kais Haddar
Sciences Faculty of Sfax
Email: kais.haddar@fss.rnu.tn

Abdelmajid Ben Hamadou
Higher Institute of Computer and Multimedia Sfax
Email: abdelmajid.benhamadou@gmail.com

The coordination is an important linguistic phenomenon. It is very frequent in various corpora and has always been a center of interest in Natural Language Processing (NLP). This importance appears mainly during the texts parsing and in the extraction of terminologies. For Arabic languages, the literature shows that researchers, working on this domain, found many problems. Indeed, the Arabic grammar is very rich and proposes several forms of coordination. In this context, our work aims to find an adequate typology of Arabic coordination and to establish an adequate grammar formalizing all its different forms. Indeed, there exist two categories of coordination: constituent coordination and non constituent coordination. The first category joins complete compounds. The second one represents cases of interaction with ellipsis phenomenon. Based on the proposed typology, we developed a Head-driven Phrase Structure grammar (HPSG) for Arabic coordination. This formalism offers a complete representation for linguistic knowledge, models grammatical principles and gives a great importance for lexical entries. Therefore, we adapted the HPSG schema to represent Arabic syntactic phenomenon and we developed two schemas to formalize the two categories of coordination. The established grammar was specified in Type Description Language (TDL). This language is designed to lexicalized grammars such as HPSG formalism. Moreover, its syntax is very similar to HPSG grammar simplifying the specification phase. Finally, to validate our grammar we used Linguistic Knowledge Building (LKB) system which is designed for grammars specified in TDL.

Interpersonal Stance and Turn-taking Behaviour in a Conversational Agent System

Merijn Bruijnes ✎
University of Twente, Human Media Interaction group
Email: m.bruijnes@utwente.nl

Teun Krikke
University of Twente, Human Media Interaction group
Email: t.f.krikke@student.utwente.nl

Rieks op den Akker
University of Twente, Human Media Interaction group
Email: h.j.a.opdenakker@utwente.nl

Embodied conversational agents have a hard time exhibiting natural, humanlike turn-taking behaviour. Most systems that have a turn taking manager (TTM) more or less implement the Sacks, Schegloff and Jefferson (SSJ) rules, which often leads to unnaturally rigid turn-taking behaviour. We feel that turn-taking is moderated by other controls, like interpersonal stance (in the sense of Leary's theory of interpersonal relationships). A recent literature review by op den Akker and Bruijnes showed that there are hardly any systems that explicitly model the relation between affect and turn-taking behaviour.

We present an affective turn-taking manager for a conversational agent that uses an internal representation of interpersonal stance to select its behaviour. The agent exhibits turn-taking based on its stance (e.g. interrupt when aggressive). The turn-taking manager is implemented using state charts. The behaviour for the agent will be based on a 'ground truth' of human-human conversations (i.e. police interrogations). Based on its stance, the agent selects and displays behaviour that is appropriate. The agent will be applied in a tutoring setting in which police trainees can learn conversational skills including proper stance taking and turn taking.

Data Services for Researchers At the Koninklijke Bibliotheek

Steven Claeyssens
Koninklijke Bibliotheek
Email: steven.claeyssens@kb.nl

The National Library of the Netherlands (KB) has planned to have digitised and OCRed its entire collection of books, periodicals and newspapers from 1470 onwards by the year 2030. But already in 2013, 10% of this enormous task will be completed, resulting in 73 million digitised pages, either from the KB itself or via public-private partnerships as Google Books and ProQuest. Many are already available via various websites (e.g. kranten.kb.nl, statengeneraaldigitaal.nl, anp.kb.nl, earlydutchbooksonline.nl) and we are working on a single entry point to (re)search all sets simultaneously.

This poster will present the full text digitised collections the KB currently has and our efforts to make these large historical Dutch text corpora available as data sets by setting up a Data Service. The service provides machine readable access to the collections and enables scholars to research entire data sets rather than just individual items.

With this poster we would like to showcase our material to the computational linguists present at CLIN 2013 and invite them to have a look at the various collections of the KB. Most of the data sets are freely available for research purposes and we welcome and encourage experiments and new applications.

Language Technology and Language Data Used in the Project IMPACT (Improving Access to Text).

Katrien Depuydt ✎
Instituut voor Nederlandse Lexicologie
Email: katrien.depuydt@inl.nl

Frank Landsbergen
Instituut voor Nederlandse Lexicologie
Email: frank.landsbergen@inl.nl

In recent years, large scale digitisation projects undertaken within cultural heritage institutions have provided access to digitised content on a scale never experienced before. While many millions of items have been made available through the internet, it represents only a small fraction of the world's cultural heritage. A major focus of the large scale digitisation initiatives has been historical texts, primarily in the form of out-of-copyright newspapers and books. However, the Optical Character Recognition (OCR) software used to translate the scanned images to machine-readable text does not provide satisfactory results for historical material. This is due to historic fonts, complex layouts, ink shining through, historical spelling variants and many other problems. In addition, a lack of institutional knowledge and expertise often causes inefficiency and 're-inventing the wheel'. The poster will detail the work of IMPACT, a large-scale integrating project funded by the European Commission as part of the Seventh Framework Programme (FP7) which aimed to improve access to historical text by innovating OCR software and language technology. Focus of the poster will be on the role of language technology and historical language data in the project. The creation of a sustainable IMPACT Centre of Competence in Text Digitisation in December 2011 will allow cultural heritage and research institutions to work together in an innovative way to continue to improve access to historical texts.

More information can be found on www.digitisation.eu.

Relations Based Summarization in "How-to" Questions Answering

Mikhail Dykov ✎
Dept. CAD/CAE Systems, Volgograd State Technical University,
Volgograd, 400131, Russia
Email: dmawork@mail.ru

Pavel Vorobkalov
Dept. CAD/CAE Systems, Volgograd State Technical University,
Volgograd, 400131, Russia
Email: pavor84@gmail.com

The problem considered in this work is searching for answers to "How-to" questions and identifying the main semantic part in the answers found. We discuss the motivation for automation of "How-to" questions answering and existing approaches. Then we describe our own approach based on document summarization using bag-of-relations. To find an answer to a specific "How-to" question we automatically retrieve documents, relevant to the question, from the results of a full-text search, then we extract grammatical relations from these documents, score them using developed method, identifying sentences that correspond to the key relations the most, and finally formulate the answer summarizing the identified sentences. We propose method of relation score evaluation taking into account relation frequency in the retrieved documents, as well as semantic closeness to other extracted relations. We also suggest a method of estimating relevance of the document containing an answer to a "How-to" question, as well as a method of identifying main objects required for performing the actions described in the document found. The developed methods were tested on a number of common "How-to" questions. Experimental results illustrated in the paper show advantage of the proposed methods compared to other considered approaches.

Minimalist Grammars with Adjunction

Meaghan Fowlie
UCLA Linguistics
Email: mfowlie@ucla.edu

I propose a model of Adjunction in Minimalist Grammars (MGs) that preserves the intuition that Adjunction is a different sort of operation from Merge, and yet is still able to account for ordering restrictions on adjuncts.

Adjuncts appear transparent to Merge. For example, a determiner (D) selects a noun (N), as in the phrase "the cat". D can select N no matter how many adjuncts N has: "the big black siamese cat". This optionality distinguishes adjunction from Merge. However, it does not mean that all adjuncts can go in any order; e.g., adjectives have ordering restrictions: *"the black siamese big cat" is degraded.

Minimalist Grammars (MGs) (Stabler 1997) offer a formalisation of References:
Chomsky (1995)'s feature-driven syntactic models. Two operations Merge and Move combine and recombine constituents based on the features lexical items introduce into the derivation. Despite differences between Adjunction and regular Merge, most MGs do not differentiate the two. My approach adds the operation Adjoin and a partial order on category features.

Such MGs with Adjunction have the same expressive power as traditional MGs but are provably more succinct, because they use features to capture both order preferences and selection requirements. They also have an advantage over treating adjunction as asymmetric feature checking, as the latter cannot account for ordering restrictions.

References:
Chomsky, Noam. 1995. The Minimalist Program. MIT Press: Cambridge MA.

Stabler, Edward. 1997. Derivational Minimalism. In C. Retoré, ed. Logical Aspects of Computational Linguistics. Springer, p68-95.

UBY – A Large-Scale Unified Lexical-Semantic Resource

Iryna Gurevych
Technical University Darmstadt
Email: gurevych@ukp.informatik.tu-darmstadt.de

Judith Eckle-Kohler
Technical University Darmstadt
Email: eckle-kohler@ukp.informatik.tu-darmstadt.de

Silvana Hartmann ✎
Technical University Darmstadt
Email: hartmann@ukp.informatik.tu-darmstadt.de

Michael Matuschek
Technical University Darmstadt
Email: matuschek@ukp.informatik.tu-darmstadt.de

Christian M. Meyer
Technical University Darmstadt
Email: meyer@ukp.informatik.tu-darmstadt.de

Tri Duc Nghiem
Technical University Darmstadt
Email: nghiem@ukp.informatik.tu-darmstadt.de

We present UBY, a large-scale lexical-semantic resource combining a wide range of information from expert-constructed and collaboratively created resources for English and German. It currently contains nine resources in two languages: English WordNet, Wiktionary, Wikipedia, FrameNet and VerbNet, German Wikipedia, Wiktionary, and GermaNet, and the multilingual OmegaWiki.

The main contributions of our work can be summarised as follows. First, we define a standardised format for modelling the heterogeneous information coming from the various lexical-semantic resources (LSRs) and languages included in UBY. For this purpose, we employ the ISO standard Lexical Markup Framework and Data Categories selected from ISOCat. In this way, all types of information provided by the LSRs in UBY are easily accessible on a fine-grained level. Further, this standardised format facilitates the extension of UBY with new languages and resources. This is different from previous efforts in combining LSRs which usually targeted particular applications and thus focused on aligning specific types of information only.

Second, UBY contains nine pairwise sense alignments between resources. Through these alignments, we provide access to the complementary information for a word sense in different resources. For example, if one looks up a particular verb sense in UBY, one has simultaneous access to the sense in WordNet and to the corresponding sense in FrameNet.

Third, UBY is freely available and we have developed an easy-to-use Java API which provides unified access to all types of information contained in UBY. This facilitates the utilization of UBY for a variety of NLP tasks.

The Good, the Bad and the Implicit: Annotating Polarity

Marjan Van de Kauter
LT3, University College Ghent, Belgium
Email: marjan.vandekauter@hogent.be

Bart Desmet ✎
LT3, University College Ghent, Belgium
Email: bart.desmet@hogent.be

Véronique Hoste
LT3, University College Ghent, Belgium
Email: veronique.hoste@hogent.be

Most of the existing sentiment annotation schemes focus on the identification of subjective statements, which explicitly express an evaluation of a certain target. Subjective statements are particularly common in user-generated content such as user reviews or blogs. However, we find that these annotation schemes are insufficient for capturing all occurrences of sentiment, which is often expressed in an implicit way. This is true especially in factual text types such as newswire, where explicit sentiment expressions are rare. We therefore propose a new annotation scheme for the fine-grained analysis of explicit as well as implicit expressions of positive and negative sentiment, also called polar expressions. This scheme was applied to a corpus of economic news articles by 8 annotators. In this presentation, we discuss the annotation scheme and the results of the annotation effort, including inter annotator agreement.

LeTs Preprocess: the Multilingual LT3 Linguistic Preprocessing Toolkit

Marjan Van de Kauter ✎
marjan.vandekauter@hogent.be

Geert Coorman
geert.coorman@hogent.be

Els Lefever
els.lefever@hogent.be

Bart Desmet
bart.desmet@hogent.be

Sofie Niemegeers
sofie.niemegeers@hogent.be

Lubbert-Jan Gringhuis
lubbertjan.gringhuis@hogent.be

Lieve Macken
lieve.macken@hogent.be

Véronique Hoste
veronique.hoste@hogent.be

LT3 Language and Translation Technology Team -
University College Ghent and Ghent University, Belgium

In the IWT-Tetra funded TExSIS project, a terminology extraction system has been developed that distills mono- and multilingual terminology lists from domain-specific corpora. The currently supported languages are Dutch, English, French and German. Since the quality of the data-driven terminology extraction approach highly depends on the linguistic preprocessing of the input text, we developed robust high-performance preprocessing modules, including POS-taggers, lemmatizers and Named Entity Recognizers for those four languages.

We present the architecture of the LeTs preprocessing pipeline and describe the methods and data used to train each component. Ten-fold cross-validation results are also presented. To assess the performance of each module on different domains, we collected real-life textual data from companies covering different domains (a.o. automotive, dredging and human resources) for Dutch, English, French and German. A manually verified gold standard for this multi-domain test corpus was created. We present the accuracy of our preprocessing tools on this corpus and compare it to the performance of other existing tools.

Controlled Automatic Translation with Euroglot?

Leo Konst
Linguistic Systems BV
Email: leokonst@euroglot.nl

It will be argued that full automatic translation still is far from being possible and may never be possible. In the meantime we can build MT-programs within a controlled environment, where the user can verify the quality of the translation and can correct or change the translation easily. Euroglot? is such a program and we will discuss and demonstrate it. The program is multilingual, rule-based and uses the intelligent and intu?tive dictionnaries of Euroglot? Professional.

Effective Unsupervised Morphological Analysis and Modeling: Statistical Study for Arabic Language

Abdellah Lakhdari ✎
Computer science department, Amar Telidji university Algeria
Email: a.lakhdari@mail.lagh-univ.dz

Dr. Cherroun Hadda
Computer science department, Amar Telidji university Algeria
Email: hadda_cherroun@mail.lagh-univ.dz

In this paper we propose a totally language independent and unsupervised method for morphological analysis, where its mastery leads us to an important improvement in many natural language processing applications, such as feature selection stage in text categorization or clustering which can be held efficiently, or enhancing language modeling vocabulary coverage as an amount step for both information retrieval and automatic speech recognition ... and many other applications. This requirement is more important in morphologically rich languages like Arabic, German, Dutch or Russian ...etc. We targeted the Arabic as an ultimate flexional language to investigate our technique, which is based on the investment of statistical inference to extract the most feasible stem for a word. The used knowledge is just letter n-gram frequencies counted from a plain text corpus of modern standard Arabic containing more than 18 million words. Once the learning input analyzed, the system result is a statistical model for morphology that can disambiguate any stemming system by mentioning the most probable stem in the non-deterministic cases.

Effect of an Unsupervised Approach using Lexical Chains On Word Sense Disambiguation

Neetu Mishra
IIITA
Email: neetumishra1508@gmail.com

This paper investigates the effect of stop word removal and lexical chains on word sense disambiguation for English and Hindi language. The evaluation has been done on a manually created corpus for nouns including their sense definitions. They are taken from Hindi WordNet developed at IIT Bombay. Context has been collected from various Hindi websites like webdunia,raftarr,and hindi webkhoj.The maximum Observed precision of 63.5% precision for over 250 instances for both,when stop word removal and Lexical chaining both has been performed.

Delemmatization Strategies for Dutch

Louis Onrust
Radboud University Nijmegen
Email: L.Onrust@student.science.ru.nl

Hans van Halteren ✎
Radboud University Nijmegen
Email: hvh@let.ru.nl

In recent years, more and more attention is given to text generation, e.g. in applications such as summarization, translation and reformulation. The final stage of the generation, however, the creation of surface forms from a lemma and desired morphosyntactic properties, has not received much attention in the literature. In this paper, we describe and evaluate various strategies for what we call delemmatization for Dutch.

We study the surface form generation for nouns, with specific number, case and diminuization, for adjectives, with specific degree and conjugation, and for verbs, with all possible verb forms. We investigate three strategies. In the pure lexicon strategy, we simply look up the desired form in a lexicon, namely e-Lex. In the pure machine learning strategy, we learn transformation rules from the information in e-Lex (using the IGtree option in Timbl) and apply the rules to generate forms. In the hybrid strategy, we use the lexicon information for known words and apply the learned rules for unknown words.

We evaluate the strategies on the basis of the manually corrected part (one million words) of the syntactically annotated LASSY corpus. We let our system take the lemma and the desired morphosyntactic information from the annotation and compare the suggested form with that actually present in the corpus. As expected, the hybrid approach yields the best results, providing the correct form for about 98% of the noun, adjective and verb tokens in LASSY.

Hybrid Approach for Corpus Based Bio-ontology Development

Rivindu Perera ✎
Informatics Institute of Technology
Email: rivindu.perera@hotmail.com

Udayangi Perera
Informatics Institute of Technology
Email: udayangi@iit.ac.lk

Bio-ontology development can be considered as a next generation requirement of providing access of health informatics for public. But still attempts and approaches taken to develop such upper merged ontology are restrained by the diverse set of features associated in biological corpora. Therefore, this paper proposes new approach to develop bio-ontology without any supervision from the controlling units. Basically this approach is empowered by two processes to extract the hidden biological knowledge to be exposed with a generic schema, relation extraction process and user-query based extraction process. Relation extraction process will analyze bio-logical corpus and will then extract semantic relations which then formalized to conceptual graphs. User-query based extraction process is used to speed up and direct the relation extraction process to knowledge which has a demand from user community. Therefore, user query based extraction process is designed to extract important terms from user queries which may be formed by users to get information about diseases, drugs or other medical information. These extracted terms are then passed to the relation extraction process to search corpus looking for an interesting text snippet to be extracted. During this search Bag-Of-Word model can also be incorporated to maximize the retrieval quality. Nevertheless, extracted relations are then transformed to a conceptual graph which will represent the agent-patient relationship with a connector element.

BioScholar: Biomedical Question Answering System

Rivindu Perera ✎
Informatics Institute of Technology
Email: rivindu.perera@hotmail.com

Udayangi Perera
Informatics Institute of Technology
Email: udayangi@iit.ac.lk

Tremendous amounts of biomedical information on the web and in publicly opened repositories have raised a new kind of issue among general public. It is to access this rich collection of information and extract knowledge that they want for a simple natural query. Empirical research carried out to address this arduous issue guided us to develop BioScholar, a question answering system which is specialized in biomedical text mining. Simply, BioScholar can be invoked from natural questions such as "what is hypothermia?", "what is headache and how to prevent it ?" or complex questions such as "what are the term variations of 9-CIS-retinoic acid?". BioScholar can generate answers according to the complexity level of question from simplest answer to the most complex answer which will associate deep and structured semantic knowledge about the problem being investigated. BioScholar works in seven different units which are focused on the responsibility assigned and maintain parallel communication strategy with related units. Term extraction and query expansion unit which is termed to be the top most unit extracts terms and key words from the questions formed by users. It is also important to notice that with probabilistic approaches, novel semantic relationship processing approaches are also mingled to extract biomedical terms from user formed questions. These extracted terms are then transformed to text search unit which involves in the text extraction from the web through a web search or from local corpora using tf-idf scores.

CGN, Asking for Collaborating Cats: ISOcat, SCHEMAcat and RELcat

Ineke Schuurman
KU Leuven and Utrecht University
Email: ineke.schuurman@ccl.kuleuven.be

Menzo Windhouwer ✎
The Language Archive - DANS
Email: Menzo.Windhouwer@dans.knaw.nl

urrently ISOcat (www.isocat.org), a Data Category Registry containing definitions of linguistic concepts, contains entries for all the concepts used in the CGN tagset. ISOcat contains relatively simple concepts like /noun/, /common/, /plural/ and /diminutive/. The registry is not meant to cover complex tags like those used in CGN, e.g., N(soort,mv,dim). So with respect to CGN we are facing two problems: the introduction of complex tags and linking those to the simple concepts contained in ISOcat proper. In conjunction with ISOcat we make use of two relatively new companion registries: SCHEMAcat, a schema registry, and RELcat, a relation registry for linguistic concepts.

The structure of the complex CGN tags can be described by a context-free grammar, e.g., an ISO 14977:1996 EBNF grammar, and its terminals and non-terminals can be annotated with ISOcat data categories:

```
(* @dcr:datcat 'N' http://www.isocat.org/datcat/DC-4909 *)
tag = 'N',    '(', NTYPE,    ',', GETAL,   ',', GRAAD ')' ... ;
(* @dcr:datcat NTYPE http://www.isocat.org/datcat/DC-4908 *)
(* @dcr:datcat 'soort' http://www.isocat.org/datcat/DC-4910 *)
(* @dcr:datcat 'eigen' http://www.isocat.org/datcat/DC-4911 *)
NTYPE = 'soort' | 'eigen' ; ...\
```

This annotated schema is stored in the SCHEMAcat registry.

There are many tagsets and large scale infrastructures like CLARIN have to deal with all of them. Using the CGN EBNF schema it is possible to parse a complex CNG tag and get a parse tree annotated with the data categories. The relation registry, RELcat, then allows making crosswalks to (closely) matching data categories that can be associated with one or more of the other tagsets.

Machine Translation: Friend Or Foe? Translator Attitude, Process, Productivity and Quality in Human Translation and Post-editing Machine Translation

Lennart Tondeleir
University College Ghent
Email: lennart.tondeleir@gmail.com

Joke Daems ✎
University College Ghent
Email: joke.daems@hogent.be

Lieve Macken
University College Ghent
Email: lieve.macken@hogent.be

As the need for translation increases, the usage of machine translation (MT) increases accordingly. While many experienced translators distrust MT output, studies have shown an increase in productivity and quality when contrasting the post-editing (PE) of machine translation with human translation. We believe that a better understanding of the PE process and product is key to the general acceptance of MT+PE by the translation community.

We present some of the findings of our research on the differences between human translation and MT followed by PE. Novice translators (Master's students in translation) were asked to perform both a translation task and a post-editing task, from English into Dutch. The selected texts were newspaper articles taken from the Dutch Parallel Corpus. Translation and post-editing processes were recorded using PET (a post-editing tool developed by Wilker Aziz and Lucia Specia). The product quality was assessed both automatically and manually. As the attitude of translators towards MT is of paramount importance for the development of the translation industry, participants had to provide feedback on the translation or post-editing process after each sentence. In the feedback for human translation, participants were asked to estimate the level of difficulty and to report on any translation issues. Feedback forms for PE tasks contained questions about the quality of the MT as well as questions to assess MT output difficulties. Translators and post-editors were also asked to indicate for which language items they consulted external resources and which ones they used.

Learning to Rank Folktale Keywords

Dolf Trieschnigg ✎
University of Twente
Email: d.trieschnigg@utwente.nl

Mariët Theune
University of Twente
Email: m.theune@utwente.nl

Theo Meder
Meertens Institute
Email: theo.meder@meertens.knaw.nl

Using keywords to describe documents can be useful for a variety of purposes, ranging from information retrieval to shallow multi-document summarization. However, manually assigning keywords to documents is a laborious and expensive task. The Dutch folktale database contains over 40,000 stories, ranging from fairy tales and legends to jokes and riddles. Each story has manual metadata assigned to it, including a list of keywords. A large backlog of digitized stories is awaiting metadata assignment before they can be included in the database. In this work we explore methods to automatically extract the most important words from folktales written in Dutch. The investigated methods range from unsupervised ranking of keywords using simple statistics to various supervised classification and ranking approaches based on more sophisticated features. The output of the automated methods is compared to freely assigned manual keywords. We show that learning to rank techniques can be succesfully applied to the task of extracting keywords.

Presenting the Language Portal, a Digital Reference Grammar for Dutch and Frisian.

Ton van der Wouden ✎
Meertens Instituut and Fryske Akademy
ton.van.der.wouden@meertens.knaw.nl

Hans Bennis
Meertens Instituut
hans.bennis@meertens.knaw.nl

Geert Booij
Universiteit Leiden
g.e.booij@hum.leidenuniv.nl

Carole Tiberius
Instituut voor Nederlandse Lexicologie
Carole.Tiberius@inl.nl

Arjen Versloot
Universiteit van Amsterdam
a.p.versloot@uva.nl

Jenny Audring
Universiteit Leiden
audringje@gmail.com

Hans Broekhuis
Meertens Insitituut
hans.broekhuis@meertens.knaw.nl

Norbert Corver
Universiteit Utrecht
N.F.M.Corver@uu.nl

Crit Cremers
Universiteit Leiden
C.L.J.M.Cremers@hum.leidenuniv.nl

Roderik Dernison
Instituut voor Nederlandse Lexicologie
Roderik.Dernison@inl.nl

Siebren Dyk
Frykse Akademy
sdyk@fryske-akademy.nl

Eric Hoekstra
Frykse Akademy
ehoekstra@fryske-akademy.nl

Frank Landsbergen
Instituut voor Nederlandse Lexicologie
frank.landsbergen@inl.nl

Kathrin Linke
Meertens Instituut
kathrin.linke@meertens.knaw.nl

Marc van Oostendorp
Meertens Instituut and Universiteit Leiden
marc.vanoostendorp@gmail.com

Willem Visser
Frykse Akademy
wvisser@fryske-akademy.nl

There are two official languages spoken in the Netherlands, Dutch and Frisian. There is, however, no complete and comprehensive scientifically based description of the grammars of these two languages. This must be seen a serious defect in a period in which language is considered an important aspect of cultural identity and cultural heritage, in which a large number of people learn the languages in question as their second language, in which educated speakers have a general lack of grammatical knowledge on their native language, and in which Dutch is an important object of study in general linguistic theory and related fields of research. Het Taalportaal/the Language Portal seeks to fill this gap. It aims at the construction of a comprehensive and authoritative

scientific grammar for Dutch and Frisian in the form of a virtual language institute. Concentrating on syntax, morphology and phonology, the Language Portal is being built around an interactive knowledge base of the current grammatical knowledge of Dutch and Frisian.

An HPSG Grammar for Arabic Relative Clauses

Ines Zalila ✎
Faculty of Economics and Management of Sfax, Tunisia
Email: ines.zalila@yahoo.fr

Kais Haddar
Sciences Faculty of Sfax, Tunisia
Email: kais.haddar@fss.rnu.tn

Abdelmajid Ben Hamadou
Institute of Computer Science and Multimedia, Sfax, Tunisia
Email: abdelmajid.benhamadou@isimsf.rnu.tn

Relative clauses present a great variety of types, and a number of properties which must be respected. For Arabic language, relative clauses are subdivided on two main forms. The first form needs an antecedent. The relative ones with antecedent generally provide information about a preceding nominal phrase (the antecedent). They are considered as a modifier for the antecedent. They are called explicative relative. The second form of Arabic relatives doesn't have an antecedent. Here, the relative clauses supplement the means of the sentence and represent an essential element in the sentence which cannot be removed. They are called completive relative. In an attempt to deal with this phenomenon, we will analyze Arabic relative clauses within the framework of Head-driven Phrase Structure Grammar (HPSG). According to Pollard and Sag (1997) approach, explicative relatives are considered as head-relative-phrase. This type of head phrase guaranties that the relative clause is the top of an unbounded dependency by the inheritance of SLASH specifications; with 'binding off' of the SLASH specification occurring at an appropriate point higher in the structure (the relative clause). Also, head-relative-phrase guaranties that relative clause modifies a nominal phrase (the antecedent) by MOD feature. In addition, we treat conjunctive nouns which introduce the second form of Arabic relatives as specifier. The aim of this project is to present different forms of Arabic relative clauses and to propose an HPSG grammar for Arabic relative clauses. The established HPSG grammar is specified in Type Description Language (TDL), experimented and evaluated with LKB platform.

Separate Training for Conditional Random Fields Using Co-occurrence Rate Factorization

Zhemin Zhu ✎
CTIT Database Group, University of Twente, Enschede, The Netherlands
Email: z.zhu@utwente.nl

Djoerd Hiemstra
CTIT Database Group, University of Twente, Enschede, The Netherlands
Email: d.hiemstra@utwente.nl

Peter Apers
CTIT Database Group, University of Twente, Enschede, The Netherlands
Email: p.m.g.apers@utwente.nl

Andreas Wombacher
CTIT Database Group, University of Twente, Enschede, The Netherlands
Email: a.wombacher@utwente.nl

Conditional Random Fields (CRFs) are undirected graphical models which are well suited to many natural language processing (NLP) tasks, such part-of-speech (POS) tagging and named entity recognition (NER). The standard training method of CRFs can be very slow for large-scale applications. As an alternative to the standard training method, piecewise training divides the full graph into pieces, trains them independently, and combines the learned weights at test time. But piecewise training does not scale well in the variable cardinality. In this paper we present separate training for undirected models based on the novel Co-occurrence Rate factorization (CR-F). Separate training is a local training method without global propagation. In contrast to directed markov models such as MEMMs, separate training is unaff ected by the label bias problem even it is a local normalized method. We do experiments on two NLP tasks, i.e., POS tagging and NER. Results show that separate training (i) is unaffected by the label bias problem; (ii) reduces the training time from weeks to seconds; and (iii) obtains competitive results to the standard and piecewise training on linear-chain CRFs. Separate training is a promising technique for scaling undirected models for natural language processing tasks. More details can be found here (http://eprints.eemcs.utwente.nl/22600/).

Rule-based Grapheme to Phoneme Conversion for Nepali Text to Speech System

Anal Haque Warsi ✎
Centre for Development of Advanced Computing (C-DAC), Kolkata
Email: anal.warsi@cdac.in

Tulika Basu
Centre for Development of Advanced Computing (C-DAC), Kolkata
Email: tulika.basu@cdac.in

This paper reports a methodology of grapheme-to-phoneme (G2P) conversion for text-to-speech synthesis in Nepali. The results of the evaluation test for grapheme to phoneme system is provided and analyzed in some detail. The contribution of this paper is twofold: on the one hand, it gives an accurate picture of the state-of-the-art in the domain of G2P conversion for Nepali. On the other hand, much room is devoted to a discussion of methodological issues for this task which may help the future researchers to work in this area.

List of authors

W

Z