Atila 2011 (Antwerp, 1-2 December)
Schedule & Abstracts

(*For all regular talks we anticipate a duration of 15 minutes, with 5 minutes for questions afterwards.*)

Thursday 1 December 2011

10.00-10.25: **Arrival** (and check-in) at Hyllit Hotel Antwerp (Seminar room: *Sala Garibaldi* – at the hotel's entrance, on your left-hand. Warm drinks provided.)

10.25-10.30: **Welcome** & introduction of Afra Alishahi (by Roser Morante)

10.30-11.30: **Invited talk** (chair: Roser Morante): *A Bayesian account of the acquisition of abstract argument structure constructions* (Afra Alishahi, invited speaker with 15 min. for questions)

> Developing computational algorithms that capture the complex structure of natural languages is an open problem. In particular, learning the abstract properties of language only from usage data without built-in knowledge of language structure remains a challenge. We have developed a Bayesian model of the acquisition and use of verb argument structure from child-directed data. In our model, the general constructions of language (such as transitive and intransitive) are viewed as a probability distribution over the syntactic and semantic features, e.g., the semantic properties of the verb and its arguments, and their relative word order in an utterance. Constructions are learned through clustering similar verb usages. Language use, on the other hand, is modeled as a Bayesian prediction problem, where the missing features in a usage are predicted based on the available parts and the acquired constructions (e.g., in sentence production, the best syntactic pattern for an utterance is predicted from the available semantic information). The model can successfully learn the common constructions of language, and its behaviour shows similarities to actual child data, both in sentence production and comprehension. Moreover, the acquired knowledge of language in this model is robust yet flexible, and many general patterns of behaviour that are observed in children can be simulated and explained by this approach.

11.30-12.30: Session 1 – **Emotions** (chair: Menno van Zaanen, 3 talks)

- *Final words: classifying emotions in suicide notes* (Bart Desmet and Véronique Hoste)

> The data set for this year's i2b2 Natural Language Processing Challenge was an unusual one. 900 suicide notes, annotated with 15 emotions or speech acts: abuse, anger, blame, fear, forgiveness, guilt, happiness or peacefulness, hopefulness, hopelessness, information, instructions, love, pride, sorrow and thankfulness. The objective of the challenge was to accurately predict their presence at the sentence level.
> We developed a system that uses 15 SVM models (1 per emotion), using the combination of features that was found to perform best on a given emotion. The features represented lexical and semantic information: lemma and trigram bag of words, and features using WordNet, SentiWordNet and a list of subjectivity clues. SVM results were improved by changing the classification threshold, using bootstrap resampling to prevent overfitting.

In this talk, we discuss the usability of this approach for emotion classification, and present the results.

- *Fine-grained emotion detection in suicide notes: Multi-label classification with probability estimates* (Kim Luyckx, Frederik Vaassen, Claudia Peersman and Walter Daelemans)

  We present a system to automatically identify emotion-carrying sentences in suicide notes and to detect the specific fine-grained emotion conveyed. With this system, we competed in Track 2 of the 2011 Medical NLP Challenge (Pestian et al., 2011), where the task was to distinguish between fifteen emotion labels, from guilt, sorrow, and hopelessness to hopefulness and happiness. An additional complication was the fact that half of the sentences was left unannotated, often not for lack of emotion, but for lack of interannotator agreement.
  Since the data set could contain multiple emotions per sentence, we adapted the system to enable assigning multiple emotion labels. The thresholding system devised to produce multi-label classification relies on probability estimates returned by an SVM classifier. Emotion labels are assigned only if their probability exceeds a certain threshold and if the probability of the sentence being emotion-free is low enough. We show the advantages of a thresholding approach by comparing it to a naïve system that assigns only the most probable label to each test sentence, and to a system trained on emotion-carrying sentences only.

- *Sentiment Analysis with Pattern 2* (Tom De Smedt and Walter Daelemans)

  The latest release of the Pattern web mining package for Python (http://www.clips.ua.ac.be/pages/pattern) contains a module for sentiment analysis for Dutch and English adjectives. We present the new open source subjectivity lexicon for Dutch adjectives. The lexicon is a dictionary of 1,100 adjectives that occur frequently in online product reviews, manually annotated with polarity strength, subjectivity and intensity, for each word sense. We discuss two machine learning methods we used to automatically expand the lexicon to 5,500 words. We evaluate the lexicon by comparing it to the user-given star rating of online product reviews. For Dutch book reviews, precision is 0.77 and recall is 0.83.
  We demonstrate a current research project that applies the lexicon to Dutch political newspaper articles.

12.30-14.00: Lunch (in the Hyllit's breakfast hall)

14.00-15.00: Session 2 – **Spelling** (chair: Walter Daelemans, 3 talks)

- *Aligning divergent text versions with anagram hashing* (Martin Reynaert)

  In the framework of the CLARIN-NL project VU-DNC we curate a diachronic corpus of Dutch newspaper texts. These have been annotated for subjectivity and quotations. The 1 MW of 1950-1951 diachronic newsaper articles have been digitized by means of Optical Character Recognition. The texts have been semi-manually corrected. As a subpart of our curation these ground truth texts are aligned with the noisy OCR-output to create gold standards for OCR post-correction research.
  We describe a new token aligner designed to facilitate building OCR gold standards. Alignment is obtained by means of anagram hashing, which allows for numerical comparison of the word strings in the ground truth versus the noisy OCR version. This

robust solution bypasses the pattern matching problems which one would encounter with a symbolical alignment procedure. The intermediate column-based format used during alignment allows for easy integration of the OCR information in the final FoLiA xml format.

- *The Chatty Corpus: a Gold Standard for Dutch Chat Normalization* (Claudia Peersman, Mike Kestemont et al.)

  Recent decades have brought a rapid succession of new communication technologies, including text messages or the numerous forms of Internet communication (e.g. e-mail, blogs, social media). An obvious effect of these recent developments has been the wild proliferation of language variation in written communication, especially affecting surface phenomena such as spelling. Speakers generally consider their standard language inadequate for these new settings and adopt a 'glocal' language variety, displaying both characteristics from a global 'Internet language' as well as their local dialect. Most NLP systems, however, were not designed to cope with the intense surface variation in Internet speech and consequently fall short in the early stage of the analysis. We discuss our annotation guidelines to create a gold standard for Dutch chat normalization and propose a first approach to automatically normalize (Flemish) Dutch chat language.

- *Deceptive Language Analysis in Italian Criminal Proceedings* (Tommaso Fornaciari)

  Methods for identifying deceptive statements in language could be of great practical use in court and in other legal situations. We focused on criminal proceedings for cases of calumny and false testimony in which the Court record contains:
  1) verbatim transcriptions of testimonies collected during the hearings;
  2) a judgment which describes the events and clearly identifies the untruthful statements for which the defendant have been found guilty.

  Thanks to this information, we were able to annotate the utterances in the testimonies as false, true, or uncertain with a high degree of confidence. This corpus was used to train models able to distinguish true from false utterances using a variety of supervised machine learning algorithms, with encouraging results.

15.00-15.30: Coffee break

15.30-17.00: **Panel discussion** between Antal van den Bosch, Véronique Hoste and Walter Daelemans (chair and moderator: Kim Luyckx)

18.00-19.30: **Social activity**

19.30u-21.00u: **Evening dinner**

Friday 2 December 2011

10.00-11.00: Session 3 – **Classification** (chair: Véronique Hoste, 3 talks)

- *Attempting to improve text classification performance using Error-Correcting Output Codes* (Frederik Vaassen and Walter Daelemans)

In the past, Error-Correcting Error Codes (ECOCs) have been used successfully to improve performance in text classification tasks (Dietterich and Bakiri, 1995). We investigate whether ECOCs remain effective for difficult classification tasks with limited datasets. We compare the performance of different codeword matrices, including a codeword matrix that takes the natural subdivisions of the classification framework into account. We compare these ECOC approaches to the more common one vs. all approach and a multiclass classifier. We observe that using ECOCs doesn't guarantee an improvement in classification performance, and that any gain to be had from using an ECOC setup will depend on the difficulty of the underlying classification task.

- *Handling skewed data: Class division* (Menno van Zaanen)

Many of the datasets we use as training and testing data in machine learning classification tasks is skewed. In this case, there is (typically only) one class that occurs much more frequently than the other classes. The existence of this large class has several effects on the task, which can be seen in the evaluation.
Having skewed data leads to a majority class baseline that is, perhaps unreasonably, high, but more importantly it may also have a negative effect on the results of the classification system. The areas in the instance space that contain instances of the non-majority classes are simply overwhelmed by the areas of the majority class.
A typical solution to the effects of skewed data is to perform downsampling. In this case, instances of the majority class as removed from the dataset randomly. This reduces the influence of the majority class, but at the same time has the negative effect that potentially interesting instances are simply removed from the dataset.
Another alternative, which we propose here, is to divide the majority class into multiple classes. The system keeps track of the new classes as being part of the original majority class, which means that no information is lost. However, the effect of the skewedness of the dataset is reduced. The major advantage of this approach is that no data instances have to be thrown away.

- *Applying domain similarity to unsupervised training data construction* (Vincent van Ash and Walter Daelemans)

When resources are scarce, additional training data can be obtained by labeling raw corpora and adding the newly labeled data to the training data. This unsupervised method can also be applied when cross-domain experiments are carried out. As has been shown in previous work, it is possible to measure the degree of similarity of corpora. Test data that is more similar to the training data will lead to better performance. The underlying hypothesis of the ongoing research here presented is that data that is more similar to the test data is more suited to be added to the training data. Preliminary results indicate that this hypothesis is valid although labeling accuracy has an influential role.

11.00-11.30: Coffee break

11.30-12.30: Session 4 – **Networks** (chair: Kim Luyckx, 3 talks)

- *The Socialist Network* (Matje van de Camp)

We develop and test machine learning-based tools for the classification of personal relationships in biographical texts, and the induction of social networks from these classifications. A case study is presented based on several hundreds of biographies of

notable persons in the Dutch social movement. Our classifiers mark relations between two persons (one being the topic of a biography, the other being mentioned in this biography) as positive, neutral, or unknown, and do so at an above-baseline level. A training set centering on a historically important person is contrasted against a multi-person training set; the latter is found to produce the most robust generalization performance. Frequency-ranked predictions of positive and negative relationships predicted by the best-performing classifier, presented in the form of person-centered social networks, are scored by a domain expert; the mean average precision results indicate that our system is better in classifying and ranking positive relations (around 70% MAP) than negative relations (around 40% MAP).

- *Discovering missing Wikipedia inter-language links by means of cross-lingual WSD* (Els Lefever and Véronique Hoste)

  We present a cross-lingual link discovery tool that discovers missing Wikipedia inter-language links to corresponding pages in other languages.
  Although the framework of our approach is language-independent, we built a prototype for our application using Dutch as an input language and Spanish, Italian, English, French and German as target languages.
  The input for our system is a set of Dutch pages for a given ambiguous noun, and the output of the system is a set of links to the corresponding pages in our five target languages.
  Our link discovery application contains two submodules. In a first step all pages are retrieved that contain a translation (in our five target languages) of the ambiguous word in the page title (Greedy crawler module), whereas in a second step all corresponding pages are linked in the focus language (being Dutch in our case) and the five target languages (Cross-lingual web page linker module).
  We consider this second step as a disambiguation task and apply a cross-lingual Word Sense Disambiguation framework to determine whether two pages refer to the same content or not.

- *Coreference resolution across different genres and the special case of bridge relations* (Orphée De Clercq, Véronique Hoste and Iris Hendrickx)

  During the SoNaR project a 1 million word core corpus has been enriched with four semantic layers: named entities, coreference relations, semantic roles and spatio-temporal relations.
  This talk will focus on the efforts put into annotating the coreference layer with four relations: identity, predicative, bridge and bound. We show how SoNaR's rich diversity enabled us to experiment with an existing mention-pair resolver (Hoste, 2005; Hendrickx et al., 2008) across various text genres. Moreover, because this is the first large-scale annotation project in which bridge relations are annotated we were also able to see whether it is possible to tune this same resolver so as to handle these special relations.
  As for as the different genres are concerned, we see that training on more diverse and genre-specific information is important but that excluding poor cross-genre material does not necessarily result in better performance. Besides, the different evaluation metrics in use for coreference research today tend to contradict each other which often hampers interpretation.
  A closer analysis of the bridge relations revealed that for humans alone, this type of relation is very complex and difficult to annotate. We see that adding semantic WordNet information does improve performance on resolving bridge relations but these

improvements are modest and more unambiguous annotation is necessary to better understand this problem.

12.30-14.00: Lunch (in the Hyllit's *Gran Duca* roof top restaurant)

14.00-15.00: Session 5 – **Grammar** (chair: Lieve Macken, 3 talks)

- *Grammar induction for assistive domestic vocal interfaces* (Janneke van de Loo, Guy De Pauw and Walter Daelemans)

  People with physical impairments who have trouble operating devices manually, could greatly benefit from vocal interfaces to control devices at home (such as the TV, radio, lights, etc.). Nevertheless, the use of assistive domestic vocal interfaces by this group of people is still far from common, due to technical and practical constraints. One of the main problems is the lack of robustness of the speech recognition system to environmental noise and to idiosyncratic pronunciations related to speech pathology and/or regional variation. Another important issue is the amount of learning and adaptation required from the user, since a restrictive vocabulary and grammar are usually preprogrammed in the system.
  The ALADIN project aims to address these problems by developing a robust, self-learning domestic vocal interface that adapts to the user instead of the other way around. The vocabulary and the grammar of the system are to be learnt on the basis of a limited set of user commands and associated controls (actions). The module for unsupervised grammar induction is designed by CLiPS. One of the targeted applications is a voice controlled computer game: patience. We have compiled a small corpus of patience commands and associated moves in a number of Wizard-of-Oz experiments. This audio corpus, manually transcribed and linguistically annotated, is used to select an appropriate grammar formalism and semantic representation for the expected range of possible commands, and as input data for some initial grammar induction experiments.

- *Evaluation of Sentence Simplification* (Sander Wubben)

  In this presentation I will address the task of Sentence Simplification and the evaluation thereof. Sentence Simplification is a popular topic and recent years have seen the development of various systems. A common approach is to align sentences from Wikipedia and Simple Wikipedia and to train some sort of simplification algorithm on the sentence pairs. To determine the success such an approach, a measure of evaluation is needed. Different measures are generally used, such as BLEU and various readability scores. In this talk I will give an overview of some of these systems, and I will discuss why sometimes the evaluation used is not very convincing. In addition to that, I will show the results of a Sentence Evalution experiment we ran ourselves.

- *Degrees of entrenchment of multi-word units: variation between units and across speakers* (Véronique Verhagen)

  Cognitive linguists argue that high-frequent word strings (e.g. a cup of tea, play hide and seek) are entrenched as one unit. Corpus frequencies are generally regarded as indicative of degree of entrenchment. However, it first needs to be determined to what extent there is individual variation, and which factors influence a word string's salience. If usage frequency determines degree of entrenchment, you would expect differences between language users, as they differ in their linguistic experiences. It is not known, though, how

large these differences are. Furthermore, while usage frequency is of great significance, it is improbable that it is the only factor determining degree of entrenchment. If a word string is salient in a different manner, it need not occur very often in order to become entrenched.
Most research into entrenchment makes use of online experiments in combination with corpus analyses. In my master's thesis, I examined the usefulness of two offline methods. I conducted a Magnitude Estimation task in which 70 Dutch native speakers judged the degree to which the words in a given word combination belong together. Analyses revealed interesting discrepancies between judgements and corpus frequencies. Furthermore, participants differed substantially in perceived degree of bondedness of the same multi-word combination.
Subsequent focus group discussions revealed that part of this variation is related to variation in usage. The discussions indicate that judgements are based on an intricate interplay of features, involving absolute and relative frequency, as well as other aspects influencing an item's prototypicality and familiarity of form and meaning.
I will outline my PhD plans to compare different forms of language processing ((re)production, reception and evaluation), and different research methods (offline and online tasks and corpus research) more extensively.

15.00-15.30: Coffee break

15.30-16.10: Session 6 – **Software** (chair: Antal van den Bosch, 2 talks)

- *PyNLPl: Python Natural Language Processing Library* (Maarten Van Gompel)

  The Python Natural Language Processing Library (PyNLPl for short, pronounce as: pineapple) is a collection of reusable Python modules for varous Natural Language Processing tasks. The modules act as building blocks to facilitate the development of NLP tools. There are a number of generic modules, for instance for basic text processing, statistics, and search problems. There are also more specific modules for dealing with a variety of formats commonly found within the context of Dutch Computational Linguistics, and for interfacing with software such as Frog and Timbl.

- *Federated Search* (Herman Stehouwer)

  Within scientific institutes there exist many language resources. These resources are often quite specialized and relatively unknown. The current infrastructural initiatives try to tackle this issue by collecting metadata about the resources and establishing centers with stable repositories to ensure the availability of the resources. However, due to the heterogeneity of the data collections and the available systems, it is currently impossible to directly search over these resources in an integrated manner. It would be beneficial if the researcher could, by means of a simple query, determine which resources and which centers contain information beneficial to his or her research, or even work on a set of distributed resources as a virtual corpus. We propose an architecture for a distributed search environment. We describe a number of existing archives and corpora participating in the project. We also detail our experiences with adapting the search systems for the architecture. The described infrastructure shall provide a location for researchers to perform searches in many different language resources.

16.10-16.30: Farewell and announcements