Mini-workshop on computational linguistics issues related to machine translation, multilinguality and vector space approaches

4th of May

Battelle, room 432

Mini-workshop

14:00- 14:15 Introduction Paola Merlo

14:15-14:45 Joakim Nivre Universal Dependency Evaluation

Multilingual parser evaluation has for a long time been hampered by the lack of cross-linguistically consistent annotation. While initiatives like Universal Dependencies have greatly improved the situation, they have also raised questions about the adequacy of existing parser evaluation metrics when applied across typologically different languages. This paper argues that the usual attachment score metrics used to evaluate dependency parsers are biased in favor of analytic languages, where grammatical structure tends to be encoded in free morphemes (function words) rather than in bound morphemes (inflection). We therefore propose an alternative evaluation metric that excludes functional relations from the attachment score. We explore the effect of this change in experiments using a subset of treebanks from release v2.0 of Universal Dependencies.

14:45-15:15 Andrei Popescu-Belis: Coherence and coreference in machine translation: the case of noun phrases

The consistent translation of nouns, known as the "one translation per discourse" hypothesis, depends partially on the coreference relations between them. In this talk, I will compare two studies aiming to improve the machine translation of nouns, either with or without coreference information. In both cases, indistinct post-editing of nouns to enforce consistency appears to bring smaller benefits than flexible, learned strategies, based on local or coreference features.

15:15-15:45 Yves Scherrer Morphological Tagging for Spoken Rusyn

Rusyn is a Slavic minority language spoken predominantly in Transcarpathian Ukraine, Eastern Slovakia, and Southeastern Poland, and is most closely related to Ukrainian. I will report first results on morphological tagging of the Corpus of Spoken Rusyn, which is currently being created at the University of Freiburg. As neither annotated corpora nor parallel corpora are electronically available for Rusyn, we propose to combine existing resources from the etymologically close Slavic languages Russian, Ukrainian, Slovak, and Polish and adapt them to Rusyn. Using MarMoT as tagging toolkit, we show that a tagger trained on a balanced set of the four source languages outperforms single language taggers by about 9% absolute, and that additional knowledge acquired in an unsupervised fashion (such as automatically induced morphosyntactic lexicons or Brown clusters) lead to further improvements.

Coffee break

16:15-16:45 Jorg Tiedemann Learning from multilingual data - Using translations as semantic supervision

Translations can be seen as semantic mirrors of the original text. They make hidden properties explicit similar to what linguistic annotation does but in a rather implicit way due to language divergences and cross-lingual variations. In this talk, I will discuss ideas that exploit translations as a kind of natural supervision for learning properties such as lexico-semantic relations and non-compositionality of certain expressions. I will also sketch ideas about the use of massively parallel data sets in the task of natural language understanding.

16:45-17:15 Aurélie Herbelot High-risk learning: acquiring concepts from tiny data

Humans are able to grasp the meaning of a new word extremely rapidly: often, a single sentence suffices for an educated guess. This extraordinary ability is still out of reach for state-of-the-art computational systems. Whilst the field of distributional semantics has made much progress in modelling the meaning of words and their composition, current systems still require exposure to huge corpora to simulate basic human semantic judgments. In this talk, I'll present a neural model of nonce word acquisition which, given some previously learnt semantic knowledge, can derive a reasonable representation of a new lexical item from tiny data. The strategy used is 'high-risk' in that the system has to trust the informativeness of the provided data, and accordingly update its parameters, in a way that would normally be seen as detrimental over a large corpus. Contrarily to previously proposed methods, this approach draws on a completely standard distributional semantics architecture, opening up possibilities for creating a generic model of semantic knowledge acquisition.

17:15-17:45 James Henderson Entailment Vectors

Distributional semantics has been extremely successful at modeling semantic similarity between words. But such vector-space models have been hard to generalize to entailment. Unlike similarity, entailment is an asymmetric relation, indicating information inclusion or abstraction. In this talk I will discuss recent work on a new vector-space framework for modeling entailment. This framework includes vector operators for measuring entailment and formulas for inferring vectors from entailments. We use this framework to define a distributional semantic model which shows impressive results in predicting abstraction (hyponymy) between words.
17:45-18:00 Closing