DutchSemCor: from manual annotation to active learning

Piek Vossen[1], Attila Görög[1], Fons Laan[2], Maarten van Gompel[4], Rubén Izquierdo[3], Antal van den Bosch[4]
[1]VU University Amsterdam, De Boelelaan 1105, 1081HV Amsterdam, The Netherlands
[2]ISLA, University of Amsterdam, Science Park 904. 1098 XH Amsterdam, The Netherlands
[3]Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
[4]Radboud University Nijmegen, P.O. Box 9013, 6500 HD Nijmegen, The Netherlands

E-mail: p.vossen@let.vu.nl, a.gorog@let.vu.nl, fons.laan@gmail.com, proycon@anaproy.nl, r.izquierdo@uvt.nl, a.vandenbosch@let.ru.nl

Abstract

State of the art Word Sense Disambiguation (WSD) systems require large sense-tagged corpora to reach satisfactory results. The number of English language resources increased in the past years, while most other languages are still under-resourced. The situation is no different for Dutch. In order to overcome this data bottleneck, the DutchSemCor project will deliver a Dutch corpus that is sense-tagged with senses from the Cornetto lexical database. Part of this corpus (circa 300K examples) is manually tagged. Current statistics for the manual annotations show an average of 91% Inter-Annotator Agreement. The remainder is automatically tagged using different WSD systems (knowledge-based, supervised and a combination of these two) and validated by human annotators (active learning). The first tests show promising results: an F-score of 74.17% for supervised WSD and an F-score of 63.66% for the knowledge-based system. The project uses existing corpora compiled in other projects (SoNaR, CGN, OpenTaal); these are extended with Internet examples for word senses that are less frequent and do not (sufficiently) appear in the corpora. We developed different tools for the purpose of manual tagging and active learning (SAT) and for importing web-snippets into the corpus (Snippet-tool). We report on the status of the project, we describe the tools and the working method used for sense tagging and active learning and, finally, we show the evaluations of the WSD systems with the current training data.

Keywords: Semantic annotation; Word Sense Disambiguation; Machine Learning; Active Learning

Presentation type preference: oral