Project Proposal Machine Learning

Title project: Structured Prediction for Named Entity Recognition

Student member names & number: Joachim Daiber: 2397331
Carmen Klaussner: 2401541

Application We will obtain our training data from CoNLL-2003[2], which contains data in English and German.Additionally, we may also use the CoNLL-2002 data for Dutch and Spanish. Named Entity Recognition is a classification problem in which the goal is to correctly predict the named entity types for the tokens in a text. Table blala shows the input/output distribution.

English Data	Articles	Sentences	Tokens
Training set	946	14,987	203,621
Development set	216	3,466	51,362
Test set	231	3,684	46,435

English Data	LOC	MISC	ORG	PER
Training set	7140	3438	6321	6600
Development set	1837	922	1341	1842
Test set	1668	702	1661	1617

German Data	Articles	Sentences	Tokens
Training set	553	12,705	206,931
Development set	201	3,068	51,444
Test set	155	3,160	51,943

German Data	LOC	MISC	ORG	PER
Training set	4363	2288	2427	2773
Development set	1181	1010	1241	1401
Test set	1035	670	773	1195

Methods: We believe that Structured Prediction provides a flexible and efficient model for Named Entity Recognition. The learning algorithm Structured Perceptron with Averaging [1].

Setup of Experiments: We will compare multiple sets of features on the given training and test set, while trying to find a good balance between complexity and performance.

Chosen programming language: Python with the NumPy package

Planning: 17-23 Sep.: Data Preparation/Literature Review
24-30 Sep.: Implementation of Learning Algorithm
1-7 Okt.: Implementation of Decoding
8-14 Okt.: Improvement of Features
15-21 Okt.: Evaluation
22-28 Okt.: Paper
29-4 Nov.: Paper

References

[1] Michael Collins. Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 1–8, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[2] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.