An IR-Inspired Approach to Recovering Named Entity Tags in Broadcast News

Niraj Shrestha, Ivan Vulic, Marie-Francine Moens
Department of Computer Science, KU Leuven
Celestijnenlaan 200A, B-3001 Heverlee, Belgium
{niraj.shrestha,ivan.vulic,marie-francine.moens}@cs.kuleuven.be

Abstract

We propose a new approach to improving named entity recognition (NER) in broadcast news speech data and to adapting NER to changing name patterns in news speech data. NER refers here to the recognition of person, location and organization names. The approach proceeds in two key steps: (1) we automatically detect document alignments between highly similar speech documents and corresponding written news stories that are easily obtainable from the Web; (2) we employ term expansion techniques inspired by information retrieval methods to recover NEs that were initially missed by the speech transcriber. We have developed several models, which are able to correct wrongly transcribed NEs in the speech data, to suggest missing NEs and to correctly assign their semantic tag. The methods differ in how relevant NEs are selected from related written documents. We have downloaded 40 short broadcast news and retrieved 5532 related written news stories from Web. The automatic speech recognition system (ASR) of FBK is used to transcribe the 40 broadcast news stories. The Stanford NER system is applied on the transcribed speech data and this method forms our ***baseline NER***. The transcribed speech data misses many NEs (106 NEs missing out of the 408 NEs) and also the Stanford NER incorrectly tags many NEs (204 NE out of the 408 NEs). Our best results in NE correction and recovery are obtained when using several related written documents that jointly evidence and suggest relevant NEs. The ***intersection model*** improves the ***baseline*** result by 0.4%, 8.9%, 3.6% in terms of average precision, recall and F_1 respectively. This model recovers 31 NEs out of 106 missing NEs from the related news texts boosting the recall substantially without hurting precision. We are also able to recover many NEs (53 NEs) by boosting the recall at the expense of a slight drop in precision with a model that considers the co-occurrence statistics of a NE recognized in the speech document and one recognized in a related written document (***co-occurrence model***).

In conclusion, our experiments show that our best model improves state-of-the-art NER results on speech data especially in terms of recall while slightly improving precision.

References

1. Cao, G., Nie, J.Y., Gao, J., Robertson, S.: Selecting good expansion terms for pseudorelevance feedback. In: Proceedings of SIGIR. pp. 243–250 (2008)
2. Favre, B., Bechet, F., Nocera, P.: Robust named entity extraction from large spoken archives. In: Proceedings of EMNLP. pp. 491–498 (2005)
3. FBK ASR transcription (2013), https://hlt-tools.fbk.eu/tosca/publish/ASR/transcribe
4. Kim, M.H., Compton, P.: Improving the performance of a named entity recognition system with knowledge acquisition. In: Proceedings of EKAW. pp. 97–113 (2012)
5. Kubala, F., Schwartz, R., Stone, R., Weischedel, R.: Named entity extraction from speech. In: Proceedings of the DARPA Broadcast News Transcription and UnderstandingWorkshop. pp. 287–292 (1998)
6. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1), 3–26 (2007)
7. Stanford: Stanford NER in CoNLL 2003 (2003), http://nlp.stanford.edu/projects/project-ner.shtml