Memory-based Named Entity Recognition in Tweets

Antal van den Bosch[1] and Toine Bogers[2]

[1] Centre for Language Studies
Radboud University Nijmegen
NL-6200 HD Nijmegen, The Netherlands
a.vandenbosch@let.ru.nl
[2] Royal School of Library Information Science
Birketinget 6, DK-2300
Copenhagen, Denmark
tb@iva.dk

Abstract. We present a memory-based named entity recognition system that participated in the MSM-2013 Concept Extraction Challenge. The system expands the training set of annotated tweets with part-of-speech tags and seedlist information, and then generates a sequential memory-based tagger comprised of separate modules for known and unknown words. Two taggers are trained: one on the original capitalized data, and one on a lowercased version of the training data. The intersection of named entities in the predictions of the two taggers is kept as the final output.

1 Background

Named-entity recognition can be seen as a labeled chunking task, where all beginning and ending words of names of predefined entity categories should be correctly identified, and the category of the entity needs to be established. A well-known solution to this task is to cast it as a token-level tagging task using the IOB or BIO coding scheme [1]. Preferably, a structured learning approach is used which combines accurate token-level decisions with a more global notion of likely and syntactically correct output sequences.

Memory-based tagging [2] is a generic machine-learning-based solution to structured sequence processing that is applicable to IOB-coded chunking. The algorithm has been implemented in MBT, an open source software package.[3] MBT generates a sequential tagger that tags from left to right, taking its own previous tagging decisions into account when generating a next tag. MBT operates on two classifiers. First, the 'known words' tagger handles words in test data which it has already seen in training data, and of which it knows the potential tags. Second, the 'unknown words' tagger is invoked to tag words not seen

[3] MBT is available in Debian Science: Linguistics, http://blends.alioth.debian.org/science/tasks/linguistics and at http://ilk.uvt.nl/mbt. The software is documented in [3].

during training. Instead of the word itself it takes into account character-based features of the word, such as the last three letters and whether it is capitalized or not [2].

Named entity recognition in social media microtexts such as Twitter messages, tweets, is generally approached with regular methods, but it is also generally acknowledged that language use in tweets deviates from average written language use in various aspects: it features more spelling and capitalization variants than usual, and it may mention a larger variety of people, places and organizations than, for instance, news. Most studies report relatively low scores because of these factors [4–6].

2 System Architecture

Figure 1 displays a schematic overview of the architecture of our system. A new incoming tweet is first enriched by seed list information, that for each token in the tweet checks whether it occurs as a geographical name, or as part of a person or organization name in gazetteer lists for these three types of entities. This produces a token-level code that is either empty (`-`) or any combination of letters representing occurrence in a person name list (`P`), a geographical name list (`G`), or an organizational name list (`O`). We provide details on the resources we used in our system in Section 3. The tweet is also part-of-speech tagged by a memory-based tagger trained on the Wall Street Journal part of the Penn Treebank [7], producing Penn Treebank part-of-speech tags for all tokens at an estimated accuracy of 95.9%.

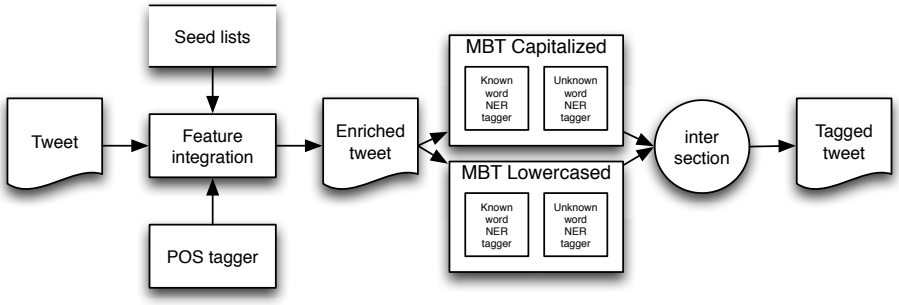

Fig. 1. The architecture of our system.

The enriched tweet is then processed by two MBT taggers. The first tagger is trained on the original training data with all capitalization information intact; the second tagger is trained on a lowercased version of the training set. The taggers both assign BIO-tags to the tokens constituting named-entity chunks [1].

The two MBT modules generate partly overlapping predictions. Only the named entity chunks that are fully identical in the output of the two modules, i.e. their intersection, are kept. The result is a tweet annotated with named entity chunks.

3 Resources

The MBT modules are trained on the official (version 1.5) training data provided for the MSM-2013 Concept Extraction Challenge.[4], complemented with the training and testing data of the CoNLL-2003 Shared Task [8] and the named-entity annotations in the ACE-2004 and ACE-2005 tasks.[5] The list of geographical names for the seedlist feature is taken from `geonames.org`;[6] Lists of person names and organization names are taken from the JRC Names corpus [9].[7].

4 Results

Table 1. Overall named entity recognition scores by the system and its components

Component	Precision	Recall	F-score
Capitalized	54.62	63.75	58.83
Lowercased	57.38	62.86	60.00
Intersection	65.82	57.21	61.21

Table 1 displays the overall scores of the final system, the intersection of the two MBT systems, together with the scores of the two systems separately. A test was run on a development set of 22,358 tokens containing 1,131 named entities extracted from the MSM-2013 training set. The capitalized MBT system attains the best recall, while the lowercased MBT attains the higher precision score. The intersection of the two predictably boosts precision at the cost of a lower recall, and attains the highest F-score of 61.21. If the gazetteer features are disabled, overall precision increases slightly from 65.8 to 66.1, but recall decreases from 57.2 to 54.9, leading to a lower F-score of 60.0. This is a predictable effect of gazetteers: they allow the recognition of more entities, but they import noise due to the context-insensitive matching of names in incorrect entity categories.

Table 2 lists the precision, recall, and F-scores on the four named entity types distinguished in the challenge. Person names are recognized more accurately than location and organization names; the miscellaneous category is hard to recognize.

[4] `http://oak.dcs.shef.ac.uk/msm2013/challenge.html`
[5] `http://projects.ldc.upenn.edu/ace/`
[6] `http://download.geonames.org/export/dump/allCountries.zip`
[7] `http://optima.jrc.it/data/entities.gzip`

Table 2. Overall named entity recognition scores on the four entity types

Named entity type	Precision	Recall	F-score
Person	75.90	69.52	72.57
Location	54.95	44.25	49.02
Organization	47.46	39.25	42.97
Miscellaneous	17.54	11.39	13.85

References

1. Tjong Kim Sang, E., Veenstra, J.: Representing text chunks. In: Proceedings of EACL'99, Bergen, Norway (1999) 173–179
2. Daelemans, W., Zavrel, J., Berck, P., Gillis, S.: MBT: A memory-based part of speech tagger generator. In Ejerhed, E., Dagan, I., eds.: Proceedings of the Fourth Workshop on Very Large Corpora, ACL SIGDAT (1996) 14–27
3. Daelemans, W., Zavrel, J., Van den Bosch, A., Van der Sloot, K.: MBT: Memory based tagger, version 3.0, reference guide. Technical Report ILK 07-04, ILK Research Group, Tilburg University (2007)
4. Ritter, A., Clark, S., Etzioni, O., et al.: Named entity recognition in tweets: an experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2011) 1524–1534
5. Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., Lee, B.S.: Twiner: Named entity recognition in targeted twitter stream. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM (2012) 721–730
6. Liu, X., Wei, F., Zhang, S., Zhou, M.: Named entity recognition for tweets. ACM Transactions on Intelligent Systems and Technology (TIST) **4**(1) (2013) 3
7. Marcus, M., Santorini, S., Marcinkiewicz, M.: Building a Large Annotated Corpus of English: the Penn Treebank. Computational Linguistics **19**(2) (1993) 313–330
8. Tjong Kim Sang, E., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Daelemans, W., Osborne, M., eds.: Proceedings of CoNLL-2003, Edmonton, Canada (2003) 142–147
9. Steinberger, R., Pouliquen, B., Kabadjov, M., Belyaeva, J., van der Goot, E.: Jrc-names: A freely available, highly multilingual named entity resource. In: Proceedings of the 8th International Conference 'Recent Advances in Natural Language Processing. (2011) 104–110