The Effect of Excluding Out of Domain Training Data from Supervised Named-Entity Recognition

Adam Persson
Department of Linguistics
Stockholm University
SE-106 91 Stockholm
adam@ling.su.se

Abstract

Supervised named-entity recognition (NER) systems perform better on text that is similar to its training data. Despite this, systems are often trained with as much data as possible, ignoring its relevance. This study explores if NER can be improved by excluding out of domain training data. A maximum entropy model is developed and evaluated twice with each domain in Stockholm-Umeå Corpus (SUC), once with all data and once with only in-domain data. For some domains, excluding out of domain training data improves tagging, but over the entire corpus it has a negative effect of less than two percentage points (both for strict and fuzzy matching).

1 Introduction

In named-entity recognition, the aim is to annotate all occurrences of explicit names, like John (person) and General Motors (organization) in a text, using some defined set of name tags. Machine learning algorithms can be trained to perform this task. However, names manifest themselves quite differently in different domains. It is challenging to create systems that perform well out of domain (Ciaramita & Altun, 2005).

In many cases, NER systems are trained with a balanced corpus to provide as much data as possible. However, this generates a very general model that perhaps is not the best possible for any one domain. This study aims to find out is such a model could be outperformed by removing all out of domain training data. This is done using a basic maximum entropy model (Berger et. al, 1996). There are of course other, more up to date methods for NER, for example various types of neural networks, such as the state of the art systems presented by Lample et al. (2016). However, the maximum entropy model is sufficient for this study as the subject of interest is the effect of excluding out of domain data, not the machine learning algorithm itself.

The results of this study have previously been presented in Persson (2016).

2 Data

Stockholm-Umeå Corpus (Källgren, 2006) is a balanced Swedish corpus, divided into nine top-level domains: reportage, editorials, reviews, skills/trades/hobbies, popular lore, biographies/essays, miscellaneous, learned/scientific, and imaginative prose. These domains differ substantially in scope and size, ranging from 17 to 127 documents per domain, each document consisting of approximately 2000 words. The distribution is inspired by The Brown Corpus (Francis & Kucera, 1964), but in SUC, imaginary prose is apparently considered a top-level domain, making it the largest one. In The Brown Corpus, imaginary prose is a section consisting of six top-level domains.

In addition to SUC, the system created for this study also takes advantage of two custom made gazetteers. One includes 1800 common Swedish person names, equal parts boys' names, girls' names, and surnames. The other one is made up of location names, including names of every Swedish town, municipality, county, and province, as well as all countries, capitals, continents, and US states. This was decided to be the very lowest level of gazetteers that any Swedish NER-system should implement.

3 System architecture

The system is implemented in Python 3.4.2 with SciKit-Learn (Pedregosa et al., 2011) 0.17.1's logistic regression (maximum entropy) model with

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 289–292,
Gothenburg, Sweden, 23-24 May 2017.

default settings, and the tagging was performed with greedy search.

Following Salomonsson et al. (2012) and Östling (2013), the SUC name tags are remapped into a more coarse-grained tag set of person, location, organization and other. On top of these name tags, the system also implements BILOU-tags (Ratinov & Roth, 2009), giving one tag to each token based on its position in a name. Multi-token names can consist of beginning (B), inside (I) and last (L), while single-token names are tagged as unit-length (U). All non-name tokens are tagged as outside (O).

The label of each observed token is a concatenation of name tag and BILOU-tag, while its features consists of the following:

- The label of the previous token.
- Word forms (3 features). 1: current token, 2: previous + current tokens, 3: current + following tokens.

  Word forms are used exactly as they appear in the text. Lemmatization or case-insensitivity is not used.
- POS-tags (2 features). 1: current token, 2: previous + following tokens.
- Matching of gazetteers (2 features). 1: person names, 2: location names.

  The gazetteer matching is done on the token level. To be able to match names in the genitive case, tokens that end with ”s” are compared twice, first with the ”s” included, and then with it removed. The same token can be matched with both gazetteers (for example Sofia, capital of Bulgaria and the corresponding girls' name).
- Pattern of capitalisation (2 features). 1: current token, 2: previous token.

  The pattern of capitalisation have five possible values: no capitalisation (xxx), full capitalisation (XXX), normal capitalisation (Xxx), sentence-initial capitalisation (. Xxx), capitalisation following dash (- Xxx), and other capitalisation (xXx, xXX, xxX).

As POS-tags are used as features, the trained system cannot be applied to raw text. It must be used as a part of a pipeline, where the text is already tokenised and POS-tagged.

These features were selected intuitively as they are some of the most common features to be used in NER-systems. Some basic testing was done during the constuction of the system, but there was no real process of structured feature selection.

4 Experiment design

To measure the effect of excluding out of domain training data, two balanced 10-fold cross-validations are carried out for each domain. The 500 documents of SUC are sorted alphanumerically with respect to their name (they are named after domain), and every tenth in-domain document is used for testing, beginning with the k'th document for each fold k. In the first cross-validation, all remaining documents in the corpus are used for training, while in the second cross-validation, only the remaining in-domain documents are used.

When the system is comparing its tagging to the gold standard for evaluation, any given name can only be part of one match, which can either be a partial match or a full match. The results of all ten folds are summed and an F1-value is calculated for the whole cross-validation.

Results are presented both for strict and fuzzy matching. Fuzzy matching accepts all names that have at least one token correctly tagged, while strict matching demands the tagging of a name to be identical to the gold standard.

In this study, the system uses SUC's gold standard POS-tagging instead of using a separate POS-tagger to prepare the test data.

5 Results

The overall result (summing all domains) shows that the in-domain training data perform slightly worse than the mixed training data. The decrease in F1-score is 1.9 percentage points for strict matching (see table 1) and 1.3 percentage points for fuzzy matching (see table 2).

There are some domains (editorial, miscellaneous) and name classes (person, institution) which are, in total, improved by excluding out of domain training data in the total count, but none of them show improvement in more than half of its domain-class combinations.

Training, tagging, and evaluating the system ninety times (10-fold cross-validation for each of the nine domains) with a 2.4 GHz Intel Xeon E5645 processor took 41 minutes using only in-

	Person	Place	Institution	Other	TOTAL
Reportage (44)	**79.3 - 78.3**	72.8 - 74.8	41.4 - 41.5	12.6 - 20.0	64.8 - 67.1
Editorial (17)	**71.4 - 70.4**	64.3 - 70.0	**66.1 - 57.6**	0.0 - 25.6	**65.3 - 64.3**
Reviews (27)	81.4 - 82.5	61.9 - 64.1	5.9 - 23.3	11.5 - 17.3	66.7 - 67.5
Skills/trades/hobbies (58)	**75.2 - 69.7**	60.6 - 68.0	46.0 - 46.7	13.2 - 20.8	56.3 - 59.9
Popular lore (48)	70.3 - 78.5	74.6 - 78.0	14.8 - 32.3	31.4 - 34.3	64.0 - 69.5
Biographies/essays (26)	78.1 - 84.3	68.3 -69.8	0.0 - 34.3	8.0 - 25.0	67.9 - 72.4
Miscellaneous (70)	**75.2 - 52.3**	78.6 - 81.4	**41.9 - 38.2**	37.9 - 40.3	**63.2 - 60.8**
Learned/scientific (83)	61.5 - 67.4	67.3 - 69.6	10.1 - 27.0	17.0 - 23.7	52.6 - 57.8
Imaginative prose (127)	68.1 - 86.6	74.4 - 74.7	0.0 - 16.5	**54.0 - 52.9**	80.3 - 80.9
TOTAL (500)	**78.5 - 78.3**	71.3 - 73.9	**39.4 - 39.0**	23.7 - 29.4	65.7 - 67.6

Table 1: Strict matching results. F1-values are presented in pairs of in-domain training data (left) and mixed training data (right). Cases where in-domain training data gets the better result are highlighted.

	Person	Place	Institution	Other	TOTAL
Reportage (44)	**84.4 - 83.3**	74.3 - 76.8	**49.3 - 48.5**	18.5 - 26.9	69.5 - 71.6
Editorial (17)	82.0 - 82.8	64.3 - 71.4	**70.8 - 61.9**	2.9 - 30.2	70.8 - 71.0
Reviews (27)	86.4 - 89.1	63.1 - 66.0	7.8 - 30.8	25.2 - 27.5	72.1 - 73.8
Skills/trades/hobbies (58)	**81.9 - 74.3**	66.5 - 70.1	49.9 - 50.9	19.2 - 27.3	61.9 - 63.9
Popular lore (48)	75.5 - 82.9	77.9 - 81.4	16.8 - 40.2	36.3 - 41.9	68.1 - 74.0
Biographies/essays (26)	84.5 - 86.5	69.0 - 72.8	0.0 - 38.7	9.6 - 28.9	72.8 - 75.1
Miscellaneous (70)	**85.6 - 56.1**	80.3 - 83.0	**55.3 - 48.2**	37.9 - 44.0	**69.7 - 65.1**
Learned/scientific (83)	68.7 - 72.0	67.7 - 70.3	12.6 - 35.7	23.5 - 29.8	57.4 - 62.0
Imaginative prose (127)	91.1 - 91.9	76.2 - 76.6	0.0 - 19.9	55.8 - 57.7	84.5 - 84.8
TOTAL (500)	**84.3 - 83.1**	73.4 - 75.8	**46.2 - 45.8**	29.3 - 35.7	70.7 - 72.0

Table 2: Fuzzy matching results. F1-values are presented in pairs of in-domain training data (left) and mixed training data (right). Cases where in-domain training data gets the better result are highlighted.

domain training data. Performing the same task with all available training data took 11 hours and 16 minutes.

6 Summary and future work

This paper describes a maximum entropy system which carries out a named-entity recognition task with different sets of training data. The purpose is to find out whether an NER-task can be improved by removing all out of domain training data for each fold in a cross-validation. Results are quite varied. Some domains and name classes show improvement, but most do not. The total count shows a worsening of less than two percentage points in both strict and fuzzy matching.

As this is a relatively small (and inconsistent) loss in performance, but a very big saving in training data size, the idea to focus more on relevance than quantity in training data should not be dismissed yet.

Future work should include normalisation of training data size, as the domains in SUC are of drastically different size. Many different training data sizes should be tested to see if there are critical points where in-domain data and mixed data stop getting better results with bigger data sets. Perhaps better results can be reached with a certain ratio of in- and out of domain training data.

On the assumption that there is a way to achieve better results by excluding (some of the) out of domain training data, an NER-system might benefit from having different models trained for different domains, and using a text-classifier to choose the appropriate model before tagging texts of unknown domain.

References

Berger, A. L., Pietra, V. J. D. & Pietra, S. A. D. 1996. *A maximum entropy approach to natural language processing.* Computational linguistics, 22(1), 39-71.

Ciaramita, M. & Altun, Y. 2005. *Named-entity recognition in novel domains with external lexical knowledge.* In Proceedings of the NIPS Workshop on Advances in Structured Learning for Text and Speech Processing.

Francis, W. N. & H. Kučera. 1964. *Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers.* Providence, Rhode Island: Department of Linguistics, Brown University. Revised 1971. Revised and amplified 1979.

Källgren, G. 2006. *Documentation of the Stockholm-Umeå Corpus.* Manual of the Stockholm Umeå Corpus version 2.0. Sofia Gustafson-Capková and Britt Hartmann (red). Stockholm University: Department of Linguistics.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. & Dyer, C. 2016. *Neural Architectures for Named Entity Recognition.* In Proceedings of NAACL-HLT, 260-270.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M. & Perrot, M. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825-2830.

Persson, A. 2016. Övervakad namntaggning med domänspecifik träningsdata. (Bachelor thesis, Stockholm University, Stockholm, Sweden) Retrieved from http://www.diva-portal.org/smash/get/diva2:934145/FULLTEXT01.pdf

Ratinov, L., & Roth, D. 2009. *Design challenges and misconceptions in named entity recognition.* Proceedings of the Thirteenth Conference on Computational Natural Language Learning, 147-155. Association for Computational Linguistics.

Salomonsson, A., Marinov, S. & Nugues, P. 2012. *Identification of entities in swedish.* SLTC 2012, 63.

Sjöbergh, J. 2003. *Combining POS-taggers for improved accuracy on Swedish text.* Proceedings of NoDaLiDa, 2003.

Östling, R. 2013. *Stagger: An open-source part of speech tagger for Swedish.* Northern European Journal of Language Technology (NEJLT), 3, 1-18.