Memory-based Classification of Proper Names in Norwegian

Anders Nøklestad

Department of Linguistics, University of Oslo
P.O. Box 1102 Blindern, N-0317 Oslo, Norway
anders.noklestad@ilf.uio.no

Abstract

This paper describes the classifier part of a named entity recogniser for Norwegian which uses memory-based learning to categorise proper names. Names are classified into one of the categories Person, Organisation, Location, Work, Event, or Other. We test the effect of using different features as input to the model, ranging from knowledge-poor features such as windows of inflected forms, to features that require high-level processing such as syntactic analysis. We run training sessions with four different *k*-values for the *k*-nearest neighbour classifier, and with four different feature weighting schemes. We also apply a document-centred approach/one sense per discourse strategy to the output of the memory-based learner. We find that the most important features are the use of gazetteers and the inclusion of lemmas that constitute multi-word proper names, and that document-centred post-processing gives a highly valuable contribution to the performance of the classifier. The best version of the classifier achieves an accuracy of 90.67% using leave-one-out testing and 83.18% using ten-fold cross-validation. The classifier outperforms a maximum entropy model using the same set of features.

1. Introduction

Automatic classification of proper names into semantic categories (Named Entity Recognition, or NER) is an important task, because it provides valuable information to other language technology tasks such as information extraction and machine translation. So far, however, work on NER for the Scandinavian languages has been scarce and limited to Swedish only (Kokkinakis 2001; Dalianis & Åström 2001). In this paper we present a named entity recogniser for Norwegian.

NER comprises two main stages: a) identification of the word or string of words that constitutes the proper name, and b) disambiguation of named entity category. This paper focuses on the second stage, presenting a named entity disambiguator or classifier which is based on memory-based learning (MBL). The first stage, which will not be our main concern here, is solved by a morphosyntactic tagger - the Oslo-Bergen tagger – which has been extended with the ability to recognise word sequences that constitute a proper name.

2. Named entity categories

The work presented here was done in the context of the Nomen Nescio network (Johannessen, 2003), and uses the set of named entity categories defined by this network:

- Person (people, animals, mythical characters, etc.)
- Organisation (companies, institutions, associations, etc.)
- Location (countries, cities, mountains, etc.)
- Work (books, movies, newspapers, etc.)
- Event (cultural events, sports events, etc.)
- Other (names that do not fit into any of the other categories)

This set extends the set of three categories (Person, Organisation, and Location) that are found in the definition of the named entity task for the two latest Message Understanding Conferences, MUC-6 and MUC-7 (Chinchor & Robinson 1998), which has been used by most previous NER work. Unlike much of the MUC-related work, we have focussed on proper names and have not attempted recognition of temporal and numerical expressions (subtasks 2 and 3 in the NE task definition for MUC), although the Oslo-Bergen tagger does in fact recognise such expressions and categorises them to a certain extent.

3. Memory-based learning

Memory-based learning (MBL) is a machine learning technique that descends from the *k*-nearest neighbour approach (see, e.g., Aha, Kibler & Albert, 1991). Other names that have been used for this kind of learning algorithm are instance-based, exemplar-based, example-based, case-based, analogical, and locally weighted learning.

Training a memory-based learner amounts to filling a database with a set of instances, each one marked with a particular category. For NLP tasks, these training instances are usually taken from a manually annotated corpus. When a new instance is to be classified, the system will search the database for the instance, or set of instances, that are most similar to the new one, i.e., its nearest neighbours in the database. Only the *k* smallest distances from the new instance are considered. The most common choice of *k* is one, but under some conditions, it is beneficial to choose higher values of *k*, notably in connection with the use of the Modified Value Distance Metric (Daelemans et al., 2002).

4. Overview of the system

4.1. The Oslo-Bergen tagger

The first component of our hybrid NER system is the Oslo-Bergen grammatical tagger (Hagen, Johannessen & Nøklestad, 2000b). The tagger performs tokenisation, part-of-speech tagging, and syntactic dependency parsing before identifying word sequences that constitute proper names. This is necessary in order to handle Norwegian proper name phrases, in which only the first word should be capitalised (e.g., Norges jeger- og fiskerforbund “The Norwegian Association of Hunters and Fishermen”), since we need NP chunking in order to delimit such sequences. See Jónsdóttir (2003) and Johannessen, Meurer & Hagen (2003) for more information about the use of the Oslo-Bergen tagger for identifying proper names.

4.2. TiMBL

For the memory-based learning part, we use the TiMBL (Tilburg Memory Based Learner) software package (Daelemans et al., 2002). We employ an IB1 database, and the Modified Value Distance Metric (MVDM) is used to obtain estimates of the match between feature values.

4.3. Document-centred post-processing

After having run a text through the Oslo-Bergen tagger and TiMBL, we test the effect of what has been referred to as a document-centred approach (Mikheev, 2000) or the concept of one sense per discourse (Yarowsky, 1995). Both of these notions refer to the phenomenon that ambiguous words have a strong tendency to occur with only one of their senses within one and the same discourse or document (Gale, Church & Yarowsky, 1992).

When classifying proper names, some contexts provide better grounds for classification than others, and using a document-centred approach (DCA), we allow the more confident decisions to override classifications made in cases where the memory-based learner is less confident.

5. Experiments

5. 1. Parameters and features

As pointed out by McDonald (1996), disambiguation of NE categories can benefit from information that is obtained from the linguistic context of the name (external evidence) as well as information that only pertains to the name itself (internal evidence). Potentially useful input features range from information that can be extracted directly from tokenised text, to information obtained from various other sources, such as gazetteers and the syntactic category of words in the sentence. Using text that has only been tokenised (i.e., not tagged or annotated with information of any kind) has the obvious advantage that the recogniser depends only on a minimum of existing NLP tools[1]. On the other hand, we would expect the use of additional information sources to increase the performance of the system.

We have run experiments with a varying set of features in order to investigate the effect of different types of information on the performance of the classifier. Each feature configuration has been tested with four different values of *k* for the *k*-nearest neighbour classifier: 5, 11, 19, and 25. For each feature combination and *k* value, we have tested the effect of using each of the feature weighting schemes provided by TiMBL: information gain, gain ratio, chi-squared and shared variance. The following types of features have been tested:

- Inflected word forms in a context window of ±2 words (IF).
- Lemmas in a context window of ±2 words (L). The lemmas are obtained by the Oslo-Bergen tagger, using a lexicon for known words and a compound analyser, or “guesser”, for unknown words.
- Whether the name consists of all capital letters (AC). The intuition behind this is that such names are likely to be acronyms, which mostly belong to the Organisation category (e.g., IBM, EU).
- Whether the name occurs in one or more gazetteer lists (GZ). The Nomen Nescio network has produced lists of names belonging to the different name categories, containing about 13,000 names in all (Person: 5,486, Location: 6,690, Organisation: 734, Work: 149, Event: 16, Other: 138).
- The number of words in the name (NW).
- Whether all the words in the name are capitalised (OC). As mentioned earlier, in proper name phrases only the first word should be capitalised. Since such phrases tend to be organisation or work names, we wanted to see whether this information would prove valuable to the classifier.
- Component lemmas (CL). When a name consists of several words, there will be one feature for each lemma of the component words.
- Suffix (SUF), taken to be the last three characters of the name.
- Syntactic relations between the name and other parts of the sentence (SYN), extracted through use of the Oslo-Bergen tagger followed by a slightly modified version of the SPARTAN system (Velldal, 2003). These relations were as follows: Subject - Verb, Object - Verb, and Preposition - Complement, where the proper name functions as Subject, Object, and Complement, respectively.
- The part-of-speech of the word to the left of the name (POS) or the two words to the left of the name (POS2). The parts-of-speech were taken from the output of the Oslo-Bergen Tagger.

5.2. Corpus

In all experiments, we use a corpus of 227,149 tokens of fiction and newspaper text, which has been extracted from the Oslo Corpus of Tagged Norwegian Texts (Hagen, Johannessen & Nøklestad, 2000a) and manually annotated with named entity categories as part of the work done within the Nomen Nescio network. The resulting corpus contains 7,537 proper names.

6. Results and evaluation

The recogniser has been trained and tested in two ways. We have used TiMBL's capacity to do leave-one-out testing (LOU), since it allows us to test each name in the corpus against all the others. We have also applied the more common method of 10-fold cross-validation (10CV). To test for statistical significance, we have applied McNemar's test to pairs of classifiers. Unless stated otherwise, the reported differences are significant at the 0.01 level.

Results from both 10CV and LOU runs are shown in Table 1. The table reports on the highest classifier accuracy that was reached for each feature combination, along with the *k* value that yielded this accuracy. As for the choice of feature weighting scheme, it turns out that gain ratio gives the best results in each case, except for the IF case with LOU, where chi-squared performs better than gain ratio, though not significantly better.

[1] However, tokenisation may benefit from other tools such as POS taggers and syntactic parsers. For example, the Oslo-Bergen tagger uses its own syntactic analysis to support tokenisation of proper names.

Features	**LOU *k*-value**	**LOU accuracy**	**10CV *k*-value**	**10CV accuracy**
IF	19	85.39	25	72.59
L	19	86.64	25	73.78
L+AC	19	86.60	25	74.05
L+GZ	11	88.40	25	79.25
L+NW	25	87.05	25	74.06
L+OC	25	87.13	25	74.38
L+CL	19	88.74	25	77.09
L+SUF	11	87.41	19	72.99
L+SYN	25	86.80	25	74.43
L+POS	19	86.92	25	74.31
L+POS2	19	86.73	25	73.99
L+ALL	5	90.66	5	81.86
L+ALL+DCA	5	90.67	5	83.18

Table 1: Results of leave-one-out testing and 10-fold cross-validation, along with optimal choices of *k*.

	Per	**Loc**	**Org**	**Work**	**Other**	**Event**
Num	3677	1912	1501	145	263	39
Recall	92.49	82.43	73.15	35.86	16.35	0.00
Prec.	88.22	77.94	73.69	76.47	43.00	0.00
F-score	90.31	80.12	73.42	48.83	23.69	0.00

Table 2 : Number of names and 10CV recall, precision and F-score for each name category.

Perhaps the most striking aspect of the figures in Table 1 is the large difference between LOU and 10CV accuracies. An advantage to using LOU is that it allows us to use the entire corpus for both training and testing, and an increased training corpus is expected to yield better results. This expectation is supported by the figures in Table 2, which shows the number of names along with recall, precision and F-score for each category when used with 10CV. The F-scores correlate strongly with the proportion of names from each category, and the few names from the Event category that are present in the corpus are in fact not recognised at all. This indicates that 10CV results might improve with a larger training corpus.

However, the use of LOU also means that when the system classifies some name, any other occurrence of the same name within the same document will be present in the instance base. With 10CV, on the other hand, we split the training and test sets on document boundaries (in order to be able to apply document-centred post-processing), so that all of the names in a certain document are either in the training set or in the test set. This is likely to be another important source of the big difference between the two result sets.

6. 1. Effects of different *k*-values

Another noticeable difference between the two training and testing schemes is that with 10CV the system performs best with the largest *k*-value in all but one of the selected-feature cases, while with LOU there is considerably more variation in the choice of optimal *k*-value. Again, this could be due to the fact that with LOU the instance base will contain more instances that are similar or identical to the test instance, meaning that the best support instances will often be found in a narrower neighbourhood.

With both schemes, the smallest *k*-value turns out to be optimal when all features are included, while the models with only selected features perform best when higher *k*-values are used. A plausible explanation for this is that a larger amount of information about the instances is likely to bring out their similarities and differences more clearly. Thus, similar instances will tend to be more tightly clustered in the instance space, and hence a small *k*-value, which restricts the search for matches to a very narrow neighbourhood, will provide the most reliable evidence for classification. However, it has been verified that *k*-values smaller than 5 do not lead to further improvement.

6. 2. Effects of individual features

In the first row of Table 1 (IF), the only features used are the inflected forms of the proper name and the words in its immediate context. Obtaining values for these features is very simple in that it only requires tokenisation of the text, and the result of this experiment can be viewed as a baseline for the performance of the MBL-based classifier.

Exchanging inflected forms for lemmas (L) yields a significant performance increase. This is to be expected, as the use of lemmas abstracts away from inflectional information which is unlikely to be important for this task, and at the same time reduces the sparseness of the data. The rest of the feature combinations all involve the use of lemmas instead of inflected forms, and further tests for statistical significance are made against the L case.

The two feature types giving the largest performance increase are component lemmas (L+CL) and gazetteers (L+GZ). A topic for further research might be to investigate whether gazetteers containing only a limited selection of names would yield comparative results, as suggested by Mikheev, Moens & Grover (1999).

Another feature leading to a smaller but still significant increase is information about whether all words of a multi-word name are capitalised (L+OC). The number of words in the name (L+NW) is significant with LOU but only borders on significance at the 0.1 level with 10CV ($p <= 0.103$). The three-letter suffix of the name (L+SUF) boosts performance significantly with LOU but actually lowers it with 10CV.

The feature type that contributes the highest level of linguistic information is the one that involves syntactic relations between names and verbs or prepositions (L+SYN). Although this feature type improves the performance of the system significantly with 10CV, with LOU it does not. In order to check whether this could be due to data sparseness, we have run additional experiments where we only included relations which occurred at least three times in the corpus. Also, since this feature type only applies to names that participate in one of the selected relations, we have compared the performance levels of the L and L+SYN classifiers on these names only. However, none of these conditions produce a significant performance increase with LOU.

Other features that use the result of linguistic processing are the part-of-speech features. L+POS gives a significant increase at the 0.05 level, while L+POS2 does not. The part-of-speech of the second word to the left receives a very low gain ratio weight value, which is a further indication that this is not a good feature for the present task.

Finally, the L and L+AC cases differ significantly only with 10CV, and then only at the 0.1 level. We suspect that this could be due to the use of newspaper texts which contain many person and location names with all capital letters (i.e., names of journalists and their locations occurring at the beginning of articles).

The document-centred post-processing step boosts performance significantly with 10CV. With LOU, on the other hand, there is virtually no increase. In this case, the instance base contains all occurrences of a name in a document except the one being tested, and this will make the classifier itself perform much the same process as DCA does.

6.3. Comparison to maximum entropy modelling

In MUC7, the best performing system was a hybrid system combining hand-written rules with a maximum entropy (MaxEnt) model (Mikheev, Moens & Grover, 1998). We were therefore interested in comparing the performance of our system to that of a MaxEnt model. As part of the work carried out within Nomen Nescio, a MaxEnt model has been trained for 200 iterations with the same set of features as the ones evaluated here, but without DCA. The present MBL model without DCA clearly outperforms the MaxEnt model, which obtains a 10CV accuracy of 76.80.

7. Conclusions

We have presented a named entity recogniser for Norwegian that reaches a high performance level and outperforms a MaxEnt model trained on the same feature set. We have investigated the contributions of different features, and have found that most of the features we tried lead to significant improvements with LOU and/or 10CV, but that the most important ones were gazetteers and the lemmas of words constituting the proper name. We also found that document-centred post-processing is a highly valuable step when used with 10CV.

8. Acknowledgements

This research was supported by the Nomen Nescio network, which is funded by the Nordic Council of Ministers through NorFA. I want to thank Janne Bondi Johannessen and Christer Johansson for their valuable comments to earlier versions of this paper.

9. References

Aha, D.W., Kibler, D., & Albert, M. (1991). Instance-based learning algorithms. Machine Learning, 6, 37--66.

Chinchor, N. & Robinson, P (1998). MUC-7 Named Entity Task Definition (version 3.5). In Proceedings of the MUC 7.

Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2002). TiMBL: Tilburg Memory-Based Learner Reference Guide. Version 4.3. ILK Technical Report. ILK 02-10.

Dalianis, H. & Åström, E. (2001). SweNam - A Swedish Named Entity recognizer. Its construction, training and evaluation. Technical report, TRITA-NA-P0113, IPLab-189. NADA, KTH.

Gale, W., Church, K., and Yarowsky, D. (1992). A Method for Disambiguating Word Senses in a Large Corpus. Computers and the Humanities, 26, 415-439.

Hagen, K., J.B. Johannessen, and A. Nøklestad (2000a). A Web-Based Advanced and User Friendly System: The Oslo Corpus of Tagged Norwegian Texts. In Gavrilidou, M., Carayannis, G., Markantonatou, S. , Piperidis, S., and Stainhaouer, G. (Eds.), Proceedings of the Second International Conference on Language Resources and Evaluation, Athens, Greece 31 May - 2 June 2000.

Hagen, K., Johannessen, J.B., & Nøklestad, A. (2000b). A Constraint- Based Tagger for Norwegian. In C.-E. Lindberg & S. Nordahl Lund (Eds.), 17th Scandinavian Conference of Linguistics, vol. I. Odense : Odense Working Papers in Language and Communication, No. 19, vol I.

Johannessen, J.B. (2003). Nomen Nescio - Nettverk for en automatisk navnegjenkjenner for norsk, svensk og dansk. In H. Holmboe (Ed.), Nordisk Sprogteknologi 2002 (pp. 327--330). Museum Tusculanums Forlag, Københavns universitet.

Johannessen, J.B., Meurer, P., & Hagen, K. (2003). Recognising word strings as names. Paper presented at Nodalida 2003, Reykjavik, Iceland.

Jónsdóttir, A.B. (2003). ARNER, what kind of name is that? *Cand. philol.* thesis. University of Oslo.

Kokkinakis, D. (2001). Design, Implementation and Evaluation of a Named-Entity Recognizer for Swedish. Research Report from the Department of Swedish, GU-ISS-01-2, Språkdata, University of Gothenburg.

McDonald, D.D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. In B. Boguraev & J. Pustejovsky (Eds.), Corpus Processing for Lexical Acquisition (pp. 21--39). Cambridge, Mass : MIT Press.

Mikheev, A. (2000). Document centered approach to text normalization. In Proceedings of SIGIR'2000 (pp. 136--143).

Mikheev, A. Grover, C., & Moens, M. (1998). Description of the LTG system used for MUC-7. In Proceedings of the 7th Message Understanding Conference (MUC-7).

Mikheev A., Moens M., & Grover C. (1999) Named Entity recognition without gazetteers. In Proceedings of the Annual Meeting of the European Association for Computational Linguistics EACL'99, Bergen, Norway (pp. 1--8).

Velldal, E. (2003). Modeling Word Senses With Fuzzy Clustering. Cand. philol. thesis. University of Oslo.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (pp. 189--196), Cambridge, MA, July 24-26. ACM Press.