SweNam-A Swedish Named Entity recognizer
Its construction, training and evaluation

Hercules Dalianis and Erik Åström
NADA-KTH
100 44 Stockholm
Sweden
Tel +46 8 790 91 05
Fax: +46 8 10 24 77
Email: {hercules,eastrom}@nada.kth.se

Abstract

In this paper we describe the development, training and evaluation of a Swedish Named Entity (NE) tagger called SweNam. NE tagging or recognition is the technique where one extracts words describing Persons, Locations, Organizations and Time from a text. We have used a combination of Machine Learning (ML) techniques and matching rules to construct our NE tagger. The training corpus consists of 108 000 Swedish news articles downloaded from Internet during 2000-2001 and we have used a number of ready NE lexicons to bootstrap our system.

The evaluation of SweNam has been carried out using 100 manually NE-annotated news texts. The results of the evaluation shows that it is possible to obtain about 92 percent precision and 46 percent recall of the named entities of a text from rule based recognition with training. In order to increase this percentage, additional training and extension of the rules and lexicons are required. A demo is available at SweNam (2001), http://www.nada.kth.se/~xmartin/swene/index-eng.html

1. Introduction

Named Entity (NE) tagging is the technique where a computer extracts those words of the texts, which represent either person, location, organization or time. This information can be used to tag a text for categorization, but can also be used to support automatic text summarization, information retrieval, topic detection tracking, etc.

2. Previous research

Automatic Named Entity recognition or tagging can be carried out with a number of methods ranging from computational linguistic methods, Stevenson & Gaizaukas (2000) to statistical and machine learning (ML) methods Bikel et al. (1997), Boisen et al. (2000), Buchholz & van den Bosch, (2000) or a mixture of both Farmakiotou et al. (2000), Mikheev et al. (1999). A nice overview of the NE area can be found in Karkaletsis et al. (1999).

The system described in Mikheev et al. (1999) was also the best performing NE-system for English in the MUC-7 (Message Understanding Conference) and had a score of 95 percent for precision and 93.6 percent for recall of Name, Organisation and Location. When also including time the score increased to 96.5 percent for precision and decreased to 88.8 percent for recall.

Some other NE-systems for English text are Identifinder from BBN (Bolt, Buranek and Newman) that is described in Boisen (2000). An early version of Identifinder called NYMBLE is described in Bikel et al. (1997). NYMBLE uses ML methods on English and Spanish texts and they found an F-score of 90 percent for Spanish texts. The same method for English gave an F-score of 93 percent. F-score is a sort of average for precision/recall. An another English system is TextPro from SRI (Stanford Research Institute, Palo Alto), which is described in (Appelt 2000).

Buchholz and van den Bosch (2000) work with Dutch texts and uses a very advanced approach by memorizing the context where known NE are found and then using this knowledge to find other NE. For some reason their results are discouraging with only 61 percent precision and 56 percent recall.

Farmakiotou et al. (2000) uses computational linguistic methods combined with ML methods, which gives pretty good results on Greek financial texts with 89.2 percent precision and 78.8 percent recall. A demo-system for Swedish for Swedish Named Entity tagging can be found at Kokkinakis (2001), though it has not been evaluated.

The definitions for Precision and Recall and for F-Score

Precision = Number of correct found entities/ number of found entities

Recall = Number of correct found entities / number of total correct entities

F-score = 2 x Recall x Precision / (Recall + Precision)

3. Approach

We have built the NE recognizer for Swedish using a mix of rules, lexicons and training strategies. All code is written in Perl version 5.6.0 (Wall et al. 1996). The strategy was to start with a reasonable small lexicons and, using rule based learning and recognition, expand these through training on large volumes of text.

This has the advantage of solving the problem of getting large enough lexicons for persons, locations and organization as these can be expanded by running the training algorithm on large enough volumes of text. (No lexicons where used for the entity time). The disadvantage is that there is always going to be a couple of incorrectly learned words.

In order to minimize the number of incorrectly learned words, the learning rules are kept very strict and only to enable learning in very ideal conditions.

As this is solely a rule based recognition approach, there will always be a number of cases where the algorithm will not recognize or even make incorrect identifications. This NE

recognizer should therefore be regarded as an experiment in how far solely rule based recognizer algorithms can go.

Rules

The rules are made up of case matching rules, which work on a sentence basis. Among these, only a few rules allow learning when identification has been done. In the case of names, a first name is only learned if it is followed by an already known last name, and there are no other words with a capital first letter next to these. This kind of very strict case learning ensures a very high precision when learning new names for the named entity lexicons.

Additionally, a minimum frequency for actual storage of the learned word further ensures high precision, as this removes most cases where the disposition of the text confuses the rule based recognizer. It is impossible to write infallible match case rules, but with this additional requirement, the word to be learned has to be recognized in more than one place. This means that the longer the training corpus is the higher the training parameter must be. The learning parameter today is greater than one (1) and has to be further tuned, depending on the length of the text (or corpus) used for learning.

For all types of named entities, a number of tests are done to remove possible incorrectly identified words. For example, the first word in every sentence is checked against a Swedish lexicon to remove non-named entity words such as regular verbs, nouns, or stop words.

The case matching rules can be summarized as follows:

Locations with a place ending of the last word, e.g. vägen (street) Sveavägen, learning of new place names (i.e matching and learning rule).

-vägen –väg, -gatan, -gata, -parken, -park, etc,

Companies and organizations

more than one capital letter in a row (Only matching rule)
e.g. AU-systems
a company ending of the last word (Matching and learning rule)
e.g. fabriken (factory) as in Framtidsfabriken)
a company type word as the last word (Matching and learning rule)
e.g. AB (Inc) as in Ericsson AB

Person names with

middle words in lower case
e.g. Hans van der Vriees
first or last name
solely lexicon based rule
title(s)
e.g. Mr Erik Åström, Vice VD Erik Åström

are only matching rules

Learning of new person names (Matching and learning rule)
- first or last name
- first, middle, or last name
  - e.g. Erik Karl Johan Åström

Time in the form (Only matching rules)
- formal date in many forms
  - e.g. den 10:e januari 2001, 10:e januari, 10 januari, 2001
- month or weekday
  - e.g. januari, måndag (January, Monday)
- date in short form
  - e.g. 2001-01-10, 1/10-2001, 1/10
- time of day
  - e.g. 10:14, 21.45

The rules are executed in the following order: Location, organization, person name and time. Each rule following another rule might make use of the previous rule's findings. We have not tested to change the order of applying the rules and what effects this can have on precision and recall.

In each of the cases of person name, company, and location, there is also a check for allowed endings and already identified words. This is done to remove possible confusing entity candidates that can be identified by their word endings or have already been identified as a different kind of entity.

In addition to checks for the other kinds of entities, there is also a check with a word endings list that contains possible confusing word endings, such as other than NE words that typically are written with a capital letter and may confuse the match case recognizer.

Nobelpriset (The Nobel Prize) is confusing for the NE since Nobel is a person but Nobelpriset is neither a person nor a organization.

Suffix lists

We use suffix lists containing allowed suffixes of locations and organizations in Swedish and prefix lists for titles for person names.

Locations
- -gatan, -området, -torget, etc

Organizations
- -firma, -byrå, -företaget, etc

Title lists
- Mr, Mrs, Miss, Herr, Fru, etc

4. The training corpus

The training is Swedish news text downloaded from 12 different news channels (See Appendix A), during 2000-2001 from Internet using an Automatic Newsagent (Hassel 2001). The corpus contains 108 000 news article with approximately 20 million words in total 40 Mb of text. To run the whole training corpus takes 12 hours on a 800 MHz PC.

5. The lexicons

The lexicons contain Swedish last and first names, locations from the whole world and Swedish organizations. The Swedish last names have been compiled from various sources (Hassel 2001). The seed list contains the 100 most common Swedish last names and variations on them. We did not use any Swedish first names in the seed list since there are more variations in first names and we need some sort of correlation between first and last names, so the learning algorithm will no make larger learning errors. The locations are taken from Gazetteers (2001). Organizations are taken from Bitweb (2001) which contains mainly Swedish companies and organizations.

Table 1. The number of words in lexicons before and after training. Before training is the same as the seed list.

Words in lexicon	First Names	Last Names	Locations	Organizations
Before Training	0	526	4268	927
After Training	741	3821	4268	3503

6. Evaluation

Precision and recall is used to evaluate the result on the test corpora.The test corpus consists of 100 manually tagged texts for name, location, organisation and time. Partial precision and recall also includes cases when only part of a NE are recognized, e.g. *Erik Åström*. has been partial recognized as *Erik*. *Sudans* (*Sudan's*) is partial recognized as *Sudan*. *TV 4 Göteborg* is partial recognized as *TV* and *17 juni* (*June 17*) is partial recognized as *june*.

The results show that training increases recall heavily on name but decreases the precision. It seems that the learning rules learn also wrong names. Many times names are confused with organization names. Both persons and organizations can also be ambiguous.

It seems that we get low recall on names and organizations due to that our system has only Swedish names and organizations in the seed list. This is to some extent true about locations, as the location lexicon even though it contains a number of international locations focuses mainly on Swedish ones.

For both company and locations, the entity ending list only contains Swedish variants of company and location endings. This is easily extended to English, and possibly other languages as well and will further increase the recall of these entities.

Person name uses location and organisation to check for no persons and location and organisation is not complete and can give errors to the name recognition.

Table 2. Precision, Recall and F-score before and after training.

Before training.

Precision	Exact precision	Partial precision
Person	0.67	0.88
Location	0.86	0.93
Organization	0.71	0.79
Time	0.74	0.89
Average precision	0.74	0.87

Recall	Exact recall	Partial recall
Person	0.05	0.07
Location	0.46	0.50
Organization	0.36	0.40
Time	0.29	0.35
Average recall	0.29	0.33

F-score	Exact F-score	Partial F-score
Person	0.10	0.13
Location	0.60	0.65
Organization	0.48	0.53
Time	0.42	0.51
Average F-score	0.40	0.45

After training

Precision	Exact precision	Partial precision
Person	0.56	0.88
Location	0.86	0.93
Organization	0.74	0.83
Time	0.71	0.89
Average precision	0.72	0.88

Recall	Exact recall	Partial recall
Person	0.28	0.44
Location	0.51	0.55
Organization	0.38	0.43
Time	0.28	0.34
Average recall	0.36	0.44

F-score	Exact F-score	Partial F-score
Person	0.38	0.59
Location	0.66	0.69
Organization	0.51	0.56
Time	0.40	0.50
Average F-score	0.49	0.59

After carrying out the first evaluation we manually controlled the results and saw that we had missed 26 entities in the manual annotation of the 100 files, in total we missed 7 persons, 4 locations and 8 organisations and 7 time equals of the originally 1774 manually annotated. Our automatical NE-tagging gave us another 1.5 percent entities which were unknown. Further on we looked into the cases where there occurred a tagging but the categorisation when wrong, that gave us another 23 tags.

Table 3. Precision, Recall and F-score after training but correlated for errors in the manual tagging and no categorisation at all.

Precision	Exact precision	Partial precision
Person	0.59	0.91
Location	0.87	0.94
Organization	0.77	0.86
Time	0.77	0.95
Average precision	0.75	**0.92**
Average precision with no special categorisation adding 23 entities	0.94	

F-score	Exact F-score	Partial F-score
Person	0.40	0.61
Location	0.67	0.70
Organization	0.53	0.58
Time	0.43	0.53
Average F-score	0.51	**0.61**
Average F-score with no special categorisation adding 23 entities	0.64	

Recall	Exact recall	Partial recall
Person	0.30	0.46
Location	0.52	0.56
Organization	0.40	0.44
Time	0.30	0.37
Average recall	0.38	**0.46**
Average recall with no special categorisation adding 23 entities		0.48

If we look at the results in Table 3. We find the SweNam gives at best 92 percent precision and 46 percent recall and a F-score of 61 percent.

7. Manual versus automatic NE tagging

We annotated manually the text corpus containing the 100 news texts which had an average length of 181 words each, i.e. 18 100 words in total. Of these we found that in the manual NE annotation we missed around 1.5 percent of the NE. These were found after we had run SweNam on the corpus giving us the not known NE's. 1.5 percent error rate on manual tagging is rather low compared to what it would have been if there were larger texts. In an example carried out at Riksdagsbiblioteket (The Swedish Parliament Library). Two human indexers where put on a task to index manual 197 rather long texts and they had only 30 percent overlap in their chosen index terms. (Bäckström 2000).

8. Conclusions and future work

This NE recognizer should be seen as a test and starting project for further work. It is in no way complete, and has to be improved in many ways in order for it be used. The aim of this project was to see how far match case recognition can go, and how it should be done.

One thing that can be improved is the amount of match case rules. Although the rules cover most of the standard forms that NE appear in, there is always a number of cases which they do not, for example to also check words in lower case around a NE candidate, e.g. *Manchester united.*

Apart from adding more rules, it would also be beneficial to expand and improve the lexicons. It is hard to learn new non-Swedish NE, if the only reliable source for recognition is solely Swedish ones. For example, there is no way for the NE recognizer to see that *Jassir Arafat* is a name because none of the words are Swedish NE and probably cannot be learned through Swedish name learning.

One other very interesting observation is that it is extremely difficult to judge if when a found NE is exact or partial correct, for example, is "Den 4:juni" (The 4:th of June) exactly the same entity as "4:juni" (4:th of June)? and "IT-säkerhetsföretaget Defcom" (The IT-security company Defcom)? is the same entity as Defcom? and what about "Zvekan-fabriken" (The Zvekan-factory)? is that the same as the company Zvekan? Where should one draw the limit? And is there any way to have a uniform way to compare results from different systems?

We plan to improve our NE tagger using better tokenization, tagging and stemming. Stemming to remove genitiv "s" in person name and organisations for example. Our NE recognizer will also be incorporated in our text summarizer SweSum (Dalianis & Hassel 2001). We are convinced that NE-recognition will give even higher performance and relevance on our automatic text summarizer SweSum (Dalianis & Hassel 2001).

The 100 manually annotated Swedish news texts with NE, keywords and Q & A, are available for these who wants to carry out research using them.

Acknowledgements

We would like to thank Martin Hassel at NADA, KTH for him supporting us with an endless stream of useful lexicons and invaluable Perl programming suggestions.

9. References

D. Appelt.(2000) TextPro (för Macintosh) June 2000 http://www.ai.sri.com/~appelt/TextPro/

D.M. Bikel, S. Miller, R.Schwartz and R. Weischedel. (1997) Nymble: a High- Performance Learning Name-finder, In Proceedings of the 5th Conference on Applied Natural Language Processing ANLP-97), Washington. D.C. pp 194-201.

Bitweb (2001) Företagsregister (Swedish Businessregister) http://www.bit.se/BIT/bol_comp.nsf/AllCompanyByName?OpenView&Start=687

S. Boisen, M.R.Crystal, R.Schwartz, R. Stone and R. Weischedel (2000) Annotating resources for Information Extraction. In the proceedings of Second International Conference on Language Resources and Evaluation. LREC-2000 Athens, Greece, 31 May- 2 June 2000, pp. 1211-1214.

S. Buchholz and A. van den Bosch (2000) Integrating seed names and ngrams for a name entity list classifier. In the proceedings of Second International Conference on Language Resources and Evaluation. LREC-2000, Athens, Greece, 31 May- 2 June 2000. pp. 1215-1221.

K. Bäckström (2000). Marknadsundersökning och utvärdering indexeringsprogram en delstudie inom projektet Automatisk indexering. (In Swedish) M.Sc. thesis. Department of Linguistics Uppsala University http://stp.ling.uu.se/educa/thesis/arch/2000-006.pdf

H. Dalianis and M. Hassel (2001): Development of a Swedish Corpus for Evaluating Summarizers and other IR-tools, Technical report, TRITA-NA-P0112, IPLab-188, NADA-KTH. http://www.nada.kth.se/~hercules/papers/TextsumEval.pdf

D. Farmakiotou, V. Karkaletsis, J. Koutsias, G. Sigletos, C.D. Spyropoulos, and P. Stamatopoulos (2000) Rule-Based Named Entity Recognition for Greek Financial texts. Comlex 2000. Proceedings of the Workshop on Computational Lexicography an Multimedia Dictionaries 22-23 September 2000, University of Patras, Greece, pp 75-78.

Gazetteers (2001) http://www.calle.com/world/

M. Hassel. (2001). Internet as Corpus - Automatic Construction of a Swedish News Corpus In the proceedings of the 13th Nordic Conference on Computational Linguistics, Uppsala May 21-22, 2001, NoDaLiDa '01.

M. Hassel. (2000). Pronominal Resolution in Automatic Text Summarisation (Master Thesis June 2000), DSV-Department of Computer and Systems Sciences, Stockholm University.

V. Karkaletsis, G. Paliouras, G. Petasis, N. Manousopoulou and C.D. Spyropoulos. (1999) Named-Entity Recognition from Greek and English Texts, Journal of Intelligent and Robotic Systems, Volume 26, No.2, pp.123-135, 1999.

G. Kokkinakis. (1998) Named Entity for Swedish. http://scrooge.spraakdata.gu.se/svedk/ne.html

A.Mikheev, C. Grover and M.Moens. (1998) Description of the LTG system used for MUC-7, http://www.ltg.ed.ac.uk/papers/muc.ps

M.Stevenson and R. Gaizaukas (2000). Using Corpus-derived Name Lists för Named Entity Recognition. Proceedings of ANLP-NAACL 6th Applied Natural Language Processing Conference and 1st Meeting of the North American Chapter of the Association for Computational Linguistics, April 29-May 4. Seattle. pp. 290-295.

SweNam (2001). Demo of SweNam http://www.nada.kth.se/~xmartin/swene/index-eng.html

L. Wall, T. Christensen, and R.L, Schwartz. (1996). *Programming Perl*. O'Reilly & Associates Inc.

Appendix A

News sources and categories used by newsAgent.

Source:	Categories:
Aftonbladet	- Economics, cultural, sports, domestic and foreign news
Amnesty International	- Press releases and news on human rights
BIT.se (Sifo Group)	- Press releases from companies
Dagens Industri	- News on the industrial market
Dagens Nyheter	- Economics. cultural. sports. domestic and foreign news
Homoplaneten (RFSL)	- News concerning rights of the homosexual community
Tidningen Mobil	- News articles on mobile communication
International Data Group	- News articles on computers
Medströms Förlag	- News articles on computers
Senaste Nytt.com	- News flashes (discontinued)
Svenska Dagbladet	- News flashes
Svenska Eko-nyheter	- News flashes
Sveriges Riksdag	- Press releases from the Swedish Parliament