Towards automatic recognition of product names: an exploratory study of brand names in economic texts

Kristina Nilsson[†*] **and Aisha Malmgren**[*]
[†]Computational Linguistics Group
[*]Department of Linguistics
Stockholm University
SE-106 91 Stockholm, Sweden
kristina.nilsson@ling.su.se, aishart@yahoo.com

Abstract

This paper describes the first stage of research towards automatic recognition of *brand names* (trademarks, product names and service names) in Swedish economic texts. The findings of an exploratory study of brand names in economic texts by Malmgren (2004) are summarized, and the work of compiling a corpus annotated with named entities based on these findings is described. A Named Entity Recognition experiment using transformation-based learning on this data shows that what is problematic to the annotator is difficult also to the recognizer; company names and brand names are closely connected and thus hard to disambiguate.

1 Introduction

In economic news texts, names denoting companies, trademarks, products and services are frequent. The production, sales and purchase of goods and services are often the topics of these texts, and such names must therefore be recognized by e.g., Information Extraction systems in this domain. However, existing guidelines for named entity annotation either include brand names in a wide category of artefacts (Sekine and Isahara, 1999) or completely ignore this name type (Chinchor, 1998), and methods for automatic recognition of names do not always make a distinction between e.g., company names and product names (Dalianis and Åström, 2001), or between product names and names of other types of objects in the domain (Kokkinakis, 2003).

Named Entity Recognition (NER) entails identifying and classifying named entities into predefined categories (Grishman, 2003). It is common for successful NER methods to use name lists in combination with context patterns (see e.g., (Volk and Clematide, 2001)) although there are several drawbacks of using such static knowledge sources; they are often domain- and language-specific, and their compilation and maintenance are time-consuming tasks. By using machine learning methods the need for static resources can be reduced. However, these methods require training and test data. Corpora annotated with named entities can thus be used to train and evaluate machine learning or statistical algorithms, and also for linguistic research and for evaluation of knowledge-based NER systems.

In this paper, we describe the work of compiling such data: a corpus of Swedish economic texts tagged with part of speech and named entities. In this corpus the following types of named entities have been found: **person**, **location**, **organization**, **financial index**, **miscellaneous**, and what can collectively be called *brand names*: **trademark**, **product**, and **service**.[1] The brand names found in the corpus have subsequently been analyzed within the framework of lexical cohesion (Halliday and Hasan, 1976). The analysis shows that 40 percent of the brand names can be classified through lexical cohesion within and across sentences, whereas the classification of the remaining 60 percent require analysis of collocations and background knowledge.

The focus of this paper is on the first stage of research in automatic recognition of brand names, building on an exploratory study of brand names in economic texts by Malmgren (2004). Below, an initial experiment –

[1] When the term *brand names* is used is this paper, we are referring to these three name types collectively.

S. Werner (ed.), *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, Ling@JoY 1, 2006, pp. 146–155
ISBN 952-458-771-8, ISSN 1796-1114. http://ling.joensuu.fi/lingjoy/

where a transformation-based learner is applied to Swedish NER using this corpus as training and testing data – is described; some problems of automatic classification of brand names are also discussed.

2 The corpus

Stockholm Ekonomikorpus (SEK) consists of about 2.800 documents collected from the online economy section of a Swedish daily newspaper. The documents in the corpus exist in two formats: in the original HTML format, and as raw text. A subset of 365 documents (about 122.000 tokens) has been chosen for further processing, which entails annotation of part of speech and named entities.

The part of speech annotation has been done automatically using TreeTagger, a probabilistic dependency-tree based tagger by Schmid (1994), trained on SUC, the Stockholm-Umeå Corpus (Ejerhed et al., 1992).[2]

Words and phrases that function as proper names in a wide sense have been annotated as named entities. This annotation has been done in two stages:

- Automatically, using the Learn-Filter-Apply-Forget method described in (Volk and Clematide, 2001) for the name types **person**, **location** and **organization**.
- Manually. The result of the automatic annotation was corrected, and the name types **trademark**, **product name** and **service name** were added (based on the study by Malmgren (2004)), as well as the additional name types **financial index**, and **miscellaneous**.

In total, there are 6725 names (tokens) in the corpus each annotated with one of these eight name types (see table 1, above).

The largest category is organization; there are 995 different names occurring 3486 times which denote companies and other types of organizations. This category includes a wide range of organizations, with the common denominator that they are organizations of people with a common goal, such as financial gain (companies), academic research and education (universities and institutes), or handling of public sector funds (governmental organizations), etc. The second largest category is person with 1408 occurrences of 618 person names.

Name type	Tokens	Types
Person	1408	618
Organization	3486	995
Location	1153	243
Trademark	28	20
Product	131	104
Service	212	76
Financial Index	241	39
Miscellaneous	67	47
Total	6725	2142

Table 1: Named Entities in SEK: tokens and types per name type, and total no of NEs.

The location category, the third largest with 1152 occurrences of 243 different names, is one of the most problematic, in that location names such as *Sverige* ('Sweden') can denote both the geographical and the geopolitical area (although the difference between these categories are not always clear), for example:

Geographical *... och huvudkontoret kommer att ligga i Sverige.* ('... and the head quarters will be located in Sweden.')

Geopolitical *... Det har upprört många fattiga länder och även Sverige har tryckt på för att deras makt ska öka.* ('Many poor countries are upset and Sweden too has argued for an increase in influence for them.')

For Information Extraction purposes, the second instance should be classified as an organization name, but in the current version of the corpus all geographical and geopolitical names are annotated as location.

In the SEK corpus, 371 occurrences of brand names have been found; among these occurrences, 28 denote 20 different trademarks, 131 denote 104 different product names, and 212 denote 76 different service names.

The category service includes names of e.g., news papers, television stations, and news agencies when these names refer to the service of providing information. Often there is also a company with the same name, and contextual clues are needed for disambiguation (see

[2]The TreeTagger has been trained for Swedish on SUC by David Hagstrand of the CL Group of the Linguistics Department, Stockholm University.

section 3.2, below). Names of financial services are also included in this category; this is a modification of the classification described in (Malmgren, 2004).

A large number of texts in the corpus give a fairly standardized account of the situation at stock markets around the world. In these texts, names of stock market indices such as *Dow Jones* and *CAC 40-index* are common. There are 241 occurrences of names denoting financial indices, but only 39 different types.[3] Finally, there are 67 names marked as miscellaneous, e.g., *Vita huset*, 'The White House', and *Internet*.

To our knowledge, there exist two other Swedish corpora annotated with named entities: the Stockholm-Umeå Corpus (SUC) of about a million words, manually annotated with 9 classes of named entities (Wennstedt, 1995), and the KTH news corpus which consists of about 108.000 Swedish news articles. In this corpus, 100 documents (about 18.000 tokens) have been manually annotated with four types of named entities: person names, locations, organizations, and time and date expressions (Hassel, 2001; Dalianis and Åström, 2001).

3 Brand names in Swedish economic texts

Brand names, and especially product names, differ from other name types such as person names and organization names in that they do not necessarily refer to a unique individual. Rather, brand names refer to a unique group of individuals that share the same name and that can be distinguished from other groups of individuals of the same kind, e.g., the group of individuals named *Volvo* can be distinguished from the group of individuals named *Saab*.

3.1 A note on terminology

As mentioned above, *brand names* is used as an umbrella term for trademarks, product names and service names. However, there is wide variation in the use of these terms both within linguistics and other fields such as marketing and law. The categorization here described is an attempt to capture the different functions that these names have in the corpus drawing on terminology described by linguists such as Piller (1996) and Clankie (2002) and by organizations like the International Trademark Association[4] and the World Intellectual Property Organization.[5]

Trademarks are viewed as having a broader scope than product and service names: while trademarks refer to diverse groups of tradable goods or services (e.g., the trademark *Volvo* identifies a range of vehicles), product and service names identify a class of (identical) objects within the group (e.g., *Volvo C 70* is one model of the trademark *Volvo*). Regarding the distinction between product and service names, it is mainly their identification of tangible and intangible concepts respectively (e.g., a car versus a TV show) that is the basis for our categorization.

3.2 Brand names and company names

Names denoting companies and brand names are closely connected. This close relation is especially noticeable when a part of a company name is found in a brand name[6]. An example from our corpus is the trademark *Volvo*, which has inherited the dominant part *Volvo* of the company name *Volvo Car Company*.

The fact that company names and brand names share the dominant causes difficulties in categorizing these names, particularly as the company name seldom occurs in its complete form in the corpus. A word like *Volvo* can refer either to the trademark or to the company itself, and can only be correctly interpreted through an analysis of the collocational context. The different name types described here are thus not discrete categories and pose a challenge to the annotators.

While trademarks usually consist of a dominant and therefore are especially difficult to disambiguate from company names, product and service names tend to be a combination of a dominant and a specific reference, e.g., *Volvo V/C 70*. However, not all brand names are created this way, but rather given a specific

[3] It could be argued that this category of names is superfluous, but as it is easier to merge categories than to subcategorize we decided to add this name type.

[4] International Trademark Association. http://www.inta.org

[5] World Intellectual Property Organization. http://www.wipo.int

[6] This part of a name, which can be used to denote both a company and its brands, is hereafter called the *dominant*.

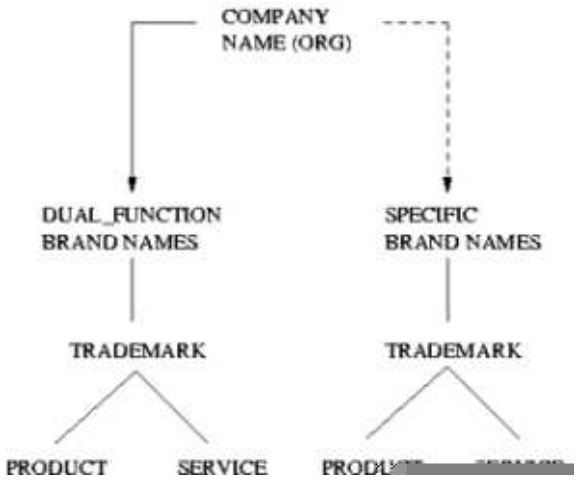

Figure 1: The relation between company names and dual-function and specific brand names.

name only. Thus, two different brand name structures can be distinguished: *dual-function brand names* that contain a dominant and a particular reference like *Volvo V/C 70*, and *specific brand names* where a dominant is not found, like *Losec*, a product of the company Astra Zeneca (see figure 1). Below are three examples of how company names and different types of brand names are used in the corpus:[7]

Company name *Volvo redovisade ett kvartalsresultat som var bättre än väntat ...* ('Volvo presented an interim result that was better than expected ...')

Trademark *Huvuddelen av omsättningen, 62 procent, kommer från de tre lastvagnsmärkena Volvo, Renault och Mack ...* ('The major part of the turnover, 62 percent, stems from the three truck brands Volvo, Renault and Mack ...')

Product name *Tätplatsen bland personbilarna har Volvo V/C 70 med 19.863 registreringar i år, ...*('Volvo V/C 70 is in the lead among passenger cars this year, with 19.863 registrations ...')

Just as company names can occur in incomplete form, so too can brand names - depending on the context. A brand name that consist of a dominant and a specific reference can occur as a partial name without the dominant and still function as a unique identifier in the context. This phenomenon can be observed in our corpus, where a mention of the dominant in the form of a company name or a trademark restricts the search space and thus allows for partial names, as in the following example where the noun phrase *företaget* ('the company') refers to the previously mentioned company name *Nokia*, allowing the partial brand name *7600*:

Partial product name *Den nya 3G-telefonen 7600, som enligt företaget ska finnas på marknaden under det fjärde kvartalet i år, är inget undantag.* ('The new 3G telephone 7600, which according to the company will be on the market during the fourth quarter this year, is no exception.')

Besides the ambiguous use of brand names, there is great variation regarding their form. Therefore, it is difficult to determine which word class these names belong to (Piller, 1996; Wennstedt, 1995). However, the trademarks, product names and service names found in the corpus tend to be treated as proper names. This is mirrored in the orthography of these names (i.e., they are capitalized) and also in their morphology, where the only processes found are the affixation of the genitive '-s' and the formation of nominal compounds.

4 Classification of brand names

To classify a name, both internal and external evidence can be used, where the internal evidence is the name itself, and external evidence is found in the context (McDonald, 1996). When examining the occurrences of brand names in the corpus, it is clear that due to the heterogeneous form of such names internal evidence is not sufficient to recognize them as denoting trademarks, products or services. Most names found in the corpus are capitalized, but capitalization is neither a prerequisite (e.g., the product name *iPod*), nor sufficient evidence for marking something as a name and, furthermore, can not be used for classification. Moreover, brand names have their origin in several different languages, e.g., *Der Spiegel* and *Firebreak*, they are made up of or resemble proper nouns, e.g., *Ericsson* and *Jack Vegas*, they display a wide range of structural variety, e.g, *AT4 CS* and *Koll.se*, and some of them are allitera-

[7] All translations are approximate.

tions or specially created words that lack meaning, e.g., *Losec*.

Due to this internal heterogeneity, name-external evidence has been analyzed by Malmgren (2004) according to the theoretical framework of textual cohesion (Halliday and Hasan, 1976). The focus has been on lexical cohesion, i.e., on the occurrences of co-referential noun phrases that allow the identification and interpretation of brand names. Through this analysis, it has been observed that about 40 percent of the occurrences of brand names found in the corpus are co-referred by common nouns located in different parts of the text, while the remaining 60 percent require analysis of collocations and background knowledge for interpretation.

4.1 Lexical cohesion within sentences

Co-references found within sentences have been analyzed as syntactic relations between the brand names and their identifications as "relations within the sentence are fairly adequately expressed already in structural terms, so that there is no need to involve the additional notion of cohesion" (Halliday and Hasan, 1976, p. 146). These relations consist mainly of appositive phrases, where the brand name is added to a common noun phrase that identifies/describes the product or service, for example:

- *... pansarvärnsroboten AT4 CS ...*('the anti-tank weapon AT4 CS')
- *... den enarmade banditen Jack Vegas ...* ('the slot machine Jack Vegas')
- *... tidskriften Der Spiegel ...* ('the magazine Der Spiegel')
- *... den nya marknadsplatsen för privatannonser Koll.se ...* ('the new portal for private ads Koll.se')

Of all brand names in the corpus, 34.8 percent can be classified on the basis of this type of lexical cohesion within sentences.

4.2 Lexical cohesion across sentences

The anaphoric co-references found across sentences can be described as reiterations, e.g., the product name *Saab 9-2* is reiterated as the definite noun phrase *bilen* ('the car'), or the generic product name *den kinesiska colan* ('the Chinese coke') is reiterated as the product name *Future Cola*. In the documents where reiterations have been found, the average number of reiterations is 4.5, which indicates that a NER system for this domain could benefit from handling this type of lexical cohesion, even though only 4.5 percent of all brand names in the corpus can be classified based on lexical cohesion across sentences.[8]

4.3 Background knowledge and collocations

Of the analyzed brand names, 60 percent lack a direct lexical identification. As a third of these names can be identified and interpreted through the collocational environment, it is possible to pattern specific constructions in which certain trademarks, product names, and service names tend to occur. However, the identification of the remaining two thirds requires background knowledge about e.g., how these types of entities typically behave, or what properties they possess. But by combining analysis of the lexical environment and assumptions about the reader's background knowledge we might be able to handle some such instances. The types of phenomenon found in the corpus include:

- Pre and post modifiers
  ... nya Mazda 6 ... ('the new Mazda 6')
  ... Delix mot högt blodtryck ... ('Delix for high blood pressure')
- Meronymic relations
  ... XC 90 med dieselmotor ... ('XC 90 with diesel engine')
- Coordination with other brand names
  ... Vauxhall Corsa, Peugeot 206, Ford Fiesta ...

It can be concluded from our corpus study that brand names that tend to occur without lexical cohesion can be assumed to be well-known to most readers, and further that these names often share the dominant with the company that produces them, e.g., *Volvo XC 90* shares the dominant *Volvo* with the producer, and that the name of the producer

[8] Unfortunately, this is a hen and egg problem: for co-reference resolution we need the named entities, and for named entity recognition we need the co-references.

can be found in the lexical environment (see section 3). In addition, the following variation in the collocational environment has been found:

Trademark is the smallest category of names with only 28 occurrences in the corpus, and the most difficult category for the human annotators. However, about a third of the instances occur with the appositive phrase *varumärket* ('the trademark').

Product names tend to co-occur with verbs like *producera* ('produce'), *sälja* ('sell'), *bygga* ('build'), *introducera* ('introduce'), *presentera* ('present'), *registrera* ('register'), *utveckla* ('develop'), and *använda* ('use'), and also with company names. Product names (P) can also be found as the complement of prepositional phrases modifying verbal nouns that are related to certain commerce domains, for example:

- *utveckling av* ('development of') + P
- *tillverkning av* ('production of') + P
- *introduktion av* ('introduction of') + P
- *försäljning av* ('sales of') + P

Service names tend to co-occur with verbs like *skriva* ('write'), *informera* ('inform'), *avslöja* ('reveal'), *rapportera* ('report'), *tillkännage* ('announce')[9], and also with company names and person names. Common constructions where service names (S) can be found are for example:

- *enligt* ('according to') + S
- *rapporterar* ('reports') + S
- *säger* ('says') + PERS + *till* ('to') + S
- *säger* + PERS + *i en intervju med* ('in an interview with') + S

A combination of background knowledge and knowledge about brand names is needed in order to classify the unknown names *XC 90*, *S 40*, and *V 40* in the example below. The background knowledge includes, e.g., that Volvo Personvagnar makes cars, and that some cars have diesel engines. The knowledge about brand names includes both knowledge about how brand names typically behave, e.g., that partial names are allowed when the dominant is shared with the company name, and that name can be found in the lexical environment, and about collocational distribution, e.g., that words such as *efterfrågan* ('demand'), *sålt* ('have sold'), *modellerna* ('the models'), and company names typically co-occur with product names.

Example *Vår ökning beror främst på stor efterfrågan på XC 90 med dieselmotor, men vi har också sålt bra av modellerna S 40 och V 40, säger Volvo Personvagnars presschef ...* ('Our increase is mostly due to great demand for XC 90 with a diesel engine, but the S 40 and V 40 models have also sold well, says the press officer of Volvo Personvagnar ...')

5 Named Entity Recognition: an initial experiment

A first experiment on supervised learning of Named Entity Recognition rules was performed using μ-TBL (Lager, 1999), a Prolog-implementation of Brill's algorithm for transformation-based learning (Brill, 1995). Named Entity Recognition entails both identification of named entities, and classification of these entities into predefined name types.

5.1 Transformation-based learning

In transformation-based learning, templates are used to learn candidate transformation rules consisting of a rewrite rule and a triggering environment from annotated training data. Learning is an iterative process, during which the output of each iteration is compared to a gold standard, and the best transformation is found. The learning continues until no transformations that will improve the annotation of the training data can be found (Brill, 1995).

Based on the analysis described in (Malmgren, 2004), 34 templates were constructed. These templates draw on three kinds of internal and external information: part of speech templates, lexical templates, and name type templates. Internal information is defined as the word itself, and any available information about part of speech. External information is found in

[9] The majority of the occurrences of service names found in the corpus denote services within communication.

the context; the words (as well as information about their part of speech) in a context window of 5 words around each word in a name. The name types of the adjacent words are also included in the definition of external information.

The part of speech templates look at the part of speech of the current word, and/or the part of speech of words in the context window; thus finding e.g., name types such as product that is often preceded by a verb and a preposition ('development of *product*').

```
Change name type tag A to B if:
  - The current word has part of speech
    tag P.
  - The preceding (following) word has
    part of speech tag P.
  - The preceding (following) two words
    have part of speech tag P and Q.
  - The word two before (after) has part
    of speech tag P.

where P and Q are variables over all
parts of speech defined in the part of
speech tag set for the corpus.
```

The lexical templates make reference to the current word and/or the adjacent words. By applying these templates to the training data, rules handling both lexical cohesion within sentences (mainly appositive phrases) and collocations can be derived.

```
Change name type tag A to B if:
  - The current word is C.
  - The preceding (following) word is C.
  - The word two before (after) is C.
  - One of the tree preceding
    (following) words is C.
  - The preceding word is C and the
    foll- owing word is D.
  - The current word or one of the two
    words before is C and one of the
    three following words are D.
  - One word in the left context window
    is C and one word in the right
    context window is D.

where C and D are variables over all
words in the training corpus.
```

The name type templates look at the name type attribute values of the adjacent words. These templates are based on the ideas that two neighboring words are likely to belong to the same type (i.e., that names are often coordinated with names of the same type) and that some name types are more likely to occur together (e.g., company names and brand names).

```
Change name type tag A to B if:
  - The preceding (following) word has
    name type attribute value T.
  - One word in the left (right) context
    window has name type attribute value
    T.

where T is a variable over all name type
attribute values.
```

Although the templates were based on the analysis described in section 3, we cannot fully exploit the findings of the study. Due to the learning method chosen for this experiment and the lack of coreference annotation in the training and test data, we are not able to explicitly model cohesion across sentences. But most reiterations should be handled by the lexical rule that says that a certain name should be marked as a specific name type, given that there is reliable evidence of the correct name type for another occurrence of this name. Further, we do not attempt to model background knowledge in this experiment, but rely on appositive phrases and collocations as modeled by the lexical templates.

Name type	Precision	Recall
Person	98.1 %	92.7 %
Organization	80.9 %	94.2 %
Location	83.3 %	82.4 %
Trademark	84.6 %	57.9 %
Product	89.3 %	76.3 %
Service	92.2 %	71.2 %
Financial Index	96.7 %	93.5 %
Miscellaneous	69.2 %	33.3 %
Overall	86.9 %	88.5 %

Table 2: NER results: precision and recall for each name type, and overall precision and recall.

5.2 Results

Training on about 100.000 words resulted in 130 rules, which in testing on about 20.000 words gave a result of overall precision of 86.9 percent, and overall recall of 88.5 percent (see table 2). Precision and recall were calculated on partial recognition, that is, when measuring the recognition of named entities consisting of

Error type		Error analysis			
(Incorrect classification > Correct name type)	Total no. of errors	No. of Apposit.	No. of Reiterat.	No. of Colloc.	No. of Backgr.
Organization > Trademark	8	0	1	6	1
Organization > Product	30	16	7	5	2
Location > Product	3	2	0	1	0
Organization > Service	17	4	7	2	4
Total	58	22	15	14	7

Table 3: NER analysis: total number of errors per error type, and grouped per type of classification clue (appositive phrase, reiteration, collocation and background knowledge).

more than one orthographic unit each recognized unit was counted individually.

The system performs well on the name types person and financial index regarding the precision, 98.1 present and 96.7 present respectively, while the best recall value 94.2 percent for organization.

Both precision and recall of the relatively closed class financial index (all in all 39 names, such as *Dow Jones* and *CAC 40*) is rather high, 96.7 and 93.5 percent respectively, to be compared to the category with the lowest scores: the heterogeneous category miscellaneous with 69.2 percent precision and 33.3 percent recall.

The lower precision values for organization names can be explained by the properties of organizations; on the one hand they are objects that can be bought or sold just like products, and on the other they can act much like humans – they can buy or sell products or other organizations, or even have opinions. The consequences of this can be seen in our results: the majority of all unrecognized brand names are classified as organization names (55 out of 58), and the majority of all unrecognized organization names are misclassified as brand names, location names, or person names.

The effects of the ambiguous use of geographical and geopolitical names (in the present version of the corpus annotated as location names, see section 2) can also be observed: within the group of location names that have been misclassified as organizations, a large group (19 out of 51) are names of geopolitical entities that behave like organizations.

The most interesting aspect of this experiment is the systems ability to correctly recognize brand names (i.e., the precision and recall of these name types). The results are encouraging with 89.3 percent precision and 76.3 percent recall for product, and 92.2 percent precision and 71.2 percent recall for service. The most difficult name type for the annotators, trademark, proved difficult also for the system to recognize with the lowest precision score of all brand names, 84.6 percent, and 57.9 percent recall, the second lowest recall score overall after miscellaneous.

5.3 Brand name error analysis

Error analysis shows that the vast majority of unrecognized brand names (55 out of 58) are misclassified as organization names, again confirming that what is difficult to the annotators is difficult to the recognizer, i.e., the distinction between organization names and brand names (see table 3).

Named entity classification can be viewed as a word sense disambiguation problem as the classifier chooses between a predefined set of name types (i.e., senses). Yarowsky (1995) has shown that, in most cases, when an ambiguous word occurs more than once in a discourse, the word sense is consistent within the discourse. This is not applicable to our classification problem due to the close relationship between company names and brand names.

The misclassified trademarks were homonyms with organization names (e.g, Nokia, Daewoo, Renault). These trademarks were also difficult to classify for the human annotators, even though typical (but in this small corpus however, infrequent) collocations could be found in 6 out of 8 cases, and all trademarks appeared in typical trademark contexts.

More than half of the identified but wrongly classified product names were marked by appositive phrases describing the product, and 7 out of the 17 erroneously classified service names were reiterations of service names that

had been misclassified as organization names due to unrecognized appositive phrases. This indicates that the results of this system might be improved by adding methods for handling appositive phrases; although the individual appositive phrases might not be frequent enough for the machine learning system, there are resources such as product ontologies which could be used for recognition of appositive phrases describing goods and services, e.g., the EU standard document for classification of products and services in common procurement, the Common Procurement Vocabulary (CPV),[10] which is available in 11 European languages (Union and Parliament, 2002).

6 Conclusion

In this paper, we have described the first stage of research towards automatic recognition of *brand names* (trademarks, product names and service names) in Swedish economic texts.

The task of Named Entity Recognition is twofold: the identification of a name, and the classification into different categories. The study on brand names in Swedish economic texts presented in this paper shows that while the brand names in the corpus can be identified as names by their orthography, classification requires analysis of the context. 40 percent of the brand names found in the corpus are coreferred by common nouns, but the remaining names cannot be classified through the study of lexical cohesion but through collocations and background knowledge (Malmgren, 2004).

A NER experiment using transformation-based learning on this data shows that what was problematic to the annotator was difficult also to the recognizer; company names and brand names are closely connected and thus hard to disambiguate, and the most problematic name type for the annotators, trademark, was also the most problematic for the recognizer. Identifying appositive phrases describing e.g., products also proved difficult to the recognizer, whereas humans have no difficulty in identifying such phrases. However, the results of the experiment are encouraging.

[10]The Common Procurement Vocabulary can be downloaded at http://simap.eu.int (last checked Sept. 6, 2005)

References

Eric Brill. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics*, December.

Nancy Chinchor. 1998. MUC-7 Named Entity Task Definition (version 3.5). In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Available from http://www.itl.nist.gov/ (Last checked Oct. 14, 2005.).

Shawn M. Clankie. 2002. *A Theory of Genericization on Brand Name Change*. The Edwin Mellen Press.

Hercules Dalianis and Erik Åström. 2001. Swenam - A Swedish Named Entity recognizer. Its construction, training, and evaluation. Technical Report TRITA-NA-P0113, NADA-KTH, July.

Eva Ejerhed, Gunnel Källgren, Ola Wennstedt, and Magnus Åström. 1992. The Linguistic Annotation System of the Stockholm-Umeå Corpus Project. Technical Report 33, Department of General Linguistics, University of Umeå.

Ralph Grishman. 2003. Information extraction. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, chapter 30, pages 545–559. Oxford University Press.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman.

Martin Hassel. 2001. Internet as Corpus. Automatic Construction of a Swedish News Corpus. In *NoDaLiDa'01*. NoDaLi.

Dimitrios Kokkinakis. 2003. Swedish NER in the Nomen Nescio Project. In *Nordisk Sprogteknologi - Nordic Language Technology. Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004*. Nordisk Sprogteknologisk Forskningsprogram.

Torbjörn Lager. 1999. The μ-TBL System: Logic Programming Tools for Transformation-Based Learning. In *Proceedings of the Third International Workshop on Computational Natural Language Learning (CoNLL'99)*, Bergen.

Aisha Malmgren. 2004. Brand names in text: An exploratory study of the identification and interpretation of brand names. Unpublished B.A. Thesis, Department of Linguistics, Stockholm University.

David D. McDonald. 1996. Internal and external evidence in the identification and semantic categorization of proper names. In B. Boguraev and J. Pustejovsky, editors, *Corpus processing for lexical acquisition*, pages 21–39. MIT.

Ingrid Piller. 1996. *American Automobile Names*. Die Blaue Eule.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Satoshi Sekine and Hitoshi Isahara. 1999. IREX project overview. In *Proceedings of the IREX workshop*, Tokyo, Japan.

European Union and European Parliament. 2002. Regulation (EC) No 2195/2002 of the European Parliament and of the Council of 5 November 2002 on the Common Procurement Vocabulary (CPV). *Official Journal L 340 , 16/12/2002 P. 0001 - 0562*.

Martin Volk and Simon Clematide. 2001. Learn-Filter-Apply-Forget. Mixed Approaches to Named Entity Recognition. In *Proc. of 6th International Workshop on Applications of Natural Language for Information Systems*, Madrid, Spain.

Ola Wennstedt. 1995. Annotering av namn i SUC-korpusen. In Kjartan G. Ottósson, Ruth V. Fjeld, and Arne Torp, editors, *The Nordic Languages and Modern Linguistics 9. Proceedings of the Ninth International Conference of Nordic and General Linguistics*, pages 315–324. University of Oslo, Novis forlag, January.

David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196.