A Knowledge-free Method for Capitalized Word Disambiguation

Andrei Mikheev*
Harlequin Ltd., Lismore House, 127 George Street, Edinburgh EH72 4JN, UK
mikheev@harlequin.co.uk

Abstract

In this paper we present an approach to the disambiguation of capitalized words when they are used in the positions where capitalization is expected, such as the first word in a sentence or after a period, quotes, etc.. Such words can act as proper names or can be just capitalized variants of common words. The main feature of our approach is that it uses a minimum of prebuilt resources and tries to dynamically infer the disambiguation clues from the entire document. The approach was thoroughly tested and achieved about 98.5% accuracy on unseen texts from The New York Times 1996 corpus.

1 Introduction

Disambiguation of capitalized words in mixed-case texts has hardly received much attention in the natural language processing and information retrieval communities, but in fact it plays an important role in many tasks. Capitalized words usually denote proper names – names of organizations, locations, people, artifacts, etc. – but there are also other positions in the text where capitalization is expected. Such ambiguous positions include the first word in a sentence, words in all-capitalized titles or table entries, a capitalized word after a colon or open quote, the first capitalized word in a list-entry, etc. Capitalized words in these and some other positions present a case of ambiguity – they can stand for proper names as in "*White* later said ...", or they can be just capitalized common words as in "*White* elephants are ...". Thus the disambiguation of capitalized words in the ambiguous positions leads to the identification of proper names[1] and in this paper we will use these two terms interchangeably. Note that this task, does not involve the classification of proper names into semantic categories (person, organization, location, etc.) which is the objective of the Named Entity Recognition task.

Many researchers observed that commonly used upper/lower case normalization does not necessarily help document retrieval. Church in (Church, 1995) among other simple text normalization techniques studied the effect of case normalization for different words and showed that "...sometimes case variants refer to the same thing (*hurricane* and *Hurricane*), sometimes they refer to different things (*continental* and *Continental*) and sometimes they don't refer to much of anything (e.g. *anytime* and *Anytime*)." Obviously these differences are due to the fact that some capitalized words stand for proper names (such as *Continental* – the name of an airline) and some don't.

Proper names are the main concern of the Named Entity Recognition subtask (Chinchor, 1998) of Information Extraction. There the disambiguation of the first word of a sentence (and in other ambiguous positions) is one of the central problems. For instance, the word "Black" in the sentence-initial position can stand for a person's surname but can also refer to the colour. Even in multi-word capitalized phrases the first word can belong to the rest of the phrase or can be just an external modifier. In the sentence "Daily, Mason and Partners lost their court case" it is clear that "Daily, Mason and Partners" is the name of a company. In the sentence "Unfortunately, Mason and Partners lost their court case" the name of the company does not involve the word "unfortunately", but

* Also at HCRC, University of Edinburgh

[1] This is not entirely true – adjectives derived from locations such as American, French, etc., are always written capitalized but in fact can stand for an adjective (*American president*) as well as a proper noun (*he was an American*).

the word "Daily" is just as common a word as "unfortunately".

Identification of proper names is also important in Machine Translation because normally proper names should be transliterated (i.e. phonetically translated) rather than properly (semantically) translated. In confidential texts, such as medical records, proper names must be identified and removed before making such texts available to unauthorized people. And in general, most of the tasks which involve different kinds of text analysis will benefit from the robust disambiguation of capitalized words into proper names and capitalized common words.

Despite the obvious importance of this problem, it was always considered part of larger tasks and, to the authors' knowledge, was not studied closely with full attention. In the part-of-speech tagging field, the disambiguation of capitalized words is treated similarly to the disambiguation of common words. However, as Church (1988) rightly pointed out "Proper nouns and capitalized words are particularly problematic: some capitalized words are proper nouns and some are not. Estimates from the Brown Corpus can be misleading. For example, the capitalized word "Acts" is found twice in Brown Corpus, both times as a proper noun (in a title). It would be misleading to infer from this evidence that the word "Acts" is always a proper noun." Church then proposed to include only high frequency capitalized words in the lexicon and also label words as proper nouns if they are "adjacent to" other capitalized words. For the rest of capitalized common words he suggested that a small probability of proper noun interpretation should be assumed and then one should hope that the surrounding context will help to make the right assignment. This approach is successful for some cases but, as we pointed out above, a sentence-initial capitalized word which is adjacent to other capitalized words is not necessarily a part of a proper name, and also many common nouns and plural nouns can be used as proper names (e.g. Riders) and their contextual expectations are not too different from their usual parts of speech.

In the Information Extraction field the disambiguation of capitalized words in the ambiguous positions was always tightly linked to the classification of the proper names into semantic classes such as person name, location, company name, etc. and to the resolution of coreference between the identified and classified proper names. This gave rise to the methods which aim at these tasks simultaneously. (Mani&MacMillan, 1995) describe a method of using contextual clues such as appositives ("PERSON, the daughter of a prominent local physician") and felicity conditions for identifying names. The contextual clues themselves are then tapped for data concerning the referents of the names. The advantage of this approach is that these contextual clues not only indicate whether a capitalized word is a proper name, but they also determine its semantic class. The disadvantage of this method is in the cost and difficulty of building a wide-coverage set of contextual clues and the dependence of these contextual clues on the domain and text genre. Contextual clues are very sensitive to the specific lexical and syntactic constructions and the clues developed for the news-wire texts are not useful for legal or medical texts.

In this paper we present a novel approach to the problem of capitalized word disambiguation. The main feature of our approach is that it uses a minimum of pre-built resources and tries to dynamically infer the disambiguation clues from the entire document under processing. This makes our approach domain and genre independent and thus inexpensive to apply when dealing with unrestricted texts. This approach was used in a named entity recognition system (Mikheev et al., 1998) where it proved to be one of the key factors in the system achieving a nearly human performance in the 7th Message Understanding Conference (MUC'7) evaluation (Chinchor, 1998).

2 Bottom-Line Performance

In general, the disambiguation of capitalized words in the mixed case texts doesn't seem to be too difficult: if a word is capitalized in an unambiguous position, e.g., not after a period or other punctuation which might require the following word to be capitalized (such as quotes or brackets), it is a proper name or part of a multi-word proper name. However, when a capitalized word is used in a position where it is expected to be capitalized, for instance, after a period or in a title, our task is to decide whether it acts

	All Words		*Known Words*		*Unknown Words*	
	tokens	*types*	*tokens*	*types*	*tokens*	*types*
Total Words	2,677	665	2,012	384	665	281
Proper Names	826	339	171	68	655	271
Common Words	1,851	326	1,841	316	10	10

Table 1: Distribution of capitalized word-tokens /word-types in the ambiguous positions.

as a proper name or as the expected capitalized common word.

The first obvious strategy for deciding whether a capitalized word in an ambiguous position is a proper name or not is to apply lexicon lookup (possibly enhanced with a morphological word guesser, e.g., (Mikheev, 1997)) and mark as proper names the words which are not listed in the lexicon of common words. Let us investigate this strategy in more detail: In our experiments we used a corpus of 100 documents (64,337 words) from The New York Times 1996. This corpus was balanced to represent different domains and was used for the formal test run of the 7th Message Understanding Conference (MUC'7) (Chinchor, 1998) in the Named Entity Recognition task.

First we ran a simple zoner which identified ambiguous positions for capitalized words – capitalized words after a period, quotes, colon, semicolon, in all-capital sentences and titles and in the beginnings of itemized list entries. The 64,337-word corpus contained 2,677 capitalized words in ambiguous positions, out of which 2,012 were listed in the lexicon of English common words. Ten common words were not listed in the lexicon and not guessed by our morphological guesser: "Forecasters", "Benchmark", "Eeverybody", "Liftoff", "Downloading", "Pretax", "Hailing", "Birdbrain", "Opting" and "Standalone". In all our experiments we did not try to disambiguate between singular and plural proper names and we also did not count as an error the adjectival reading of words which are always written capitalized (e.g. American, Russian, Okinawian, etc.). The distribution of proper names among the ambiguous capitalized words is shown in Table 1.

Table 1 allows one to estimate the performance of the lexicon lookup strategy which we take as the bottom-line. First, using this strategy we would wrongly assign the ten common words which were not listed in the lexicon. More damaging is the blind assignment of the common word category to the words listed in the lexicon: out of 2,012 known word-tokens 171 actually were used as proper names. This in total would give us 181 errors out of 2,677 tries – about a 6.76% misclassification error on capitalized word-tokens in the ambiguous positions.

The lexicon lookup strategy can be enhanced by accounting for the immediate context of the capitalized words in question. However, capitalized words in the ambiguous positions are not easily disambiguated by their surrounding part-of-speech context as attempted by part-of-speech taggers. For instance, many surnames are at the same time nouns or plural nouns in English and thus in both variants can be followed by a past tense verb. Capitalized words in the phrases *Sails rose ...* or *Feeling himself ...* can easily be interpreted either way and only knowledge of semantics disallows the plural noun interpretation of *Stars can read.*

Another challenge is to decide whether the first capitalized word belongs to the group of the following proper nouns or is an external modifier and therefore not a proper noun. For instance, *All American Bank* is a single phrase but in *All State Police* the word "All" is an external modifier and can be safely decapitalized. One might argue that a part-of-speech tagger can capture that in the first case the word "All" modified a singular proper noun ("Bank") and hence is not grammatical as an external modifier and in the second case it is a grammatical external modifier since it modifies a plural proper noun ("Police") but a simple counter-example – *All American Games* – defeats this line of reasoning.

The third challenge is of a more local nature – it reflects a capitalization convention adopted by the author. For instance, words which reflect the occupation of a person can be used in an honorific mode e.g. *"Chairman Mao"* vs.

"ATT chairman Smith" or *"Astronaut Mario Runko"* vs. *"astronaut Mario Runko"*. When such a phrase opens a sentence, looking at the sentence only, even a human classifier has troubles in making a decision.

To evaluate the performance of part-of-speech taggers on the proper-noun identification task we ran an HMM trigram tagger (Mikheev, 1997) and the Brill tagger (Brill,.1995) on our corpus. Both taggers used the Penn Treebank tagset and were trained on the Wall Street Journal corpus (Marcus et al., 1993). Since for our task the mismatch between plural proper noun (NNPS) and singular proper noun (NNP) was not important we did not count this as an error. Depending on the smoothing technique, the HMM tagger performed in the range of 5.3%-4.5% of the misclassification error on capitalized common words in the ambiguous positions, and the Brill tagger showed a similar pattern when we varied the lexicon acquisition heuristics.

The taggers handled the cases when a potential adjective was followed by a verb or adverb (*"Golden added .."*) well but they got confused with a potential noun followed by a verb or adverb (*"Butler was .."* vs. *"Safety was .."*), probably because the taggers could not distinguish between concrete and mass nouns. Not surprisingly the taggers did not do well on potential plural nouns and gerunds - none of them were assigned as a proper noun. The taggers also could not handle the case when a potential noun or adjective was followed by another capitalized word ("General Accounting Office") well. In general, when the taggers did not have strong lexical preferences, apart from several obvious cases they tended to assign a common word category to known capitalized words in the ambiguous positions and the performance of the part-of-speech tagging approach was only about 2% superior to the simple bottom-line strategy.

3 Our Knowledge-Free Method

As we discussed above, the bad news (well, not really news) is that virtually any common word can potentially act as a proper name or part of a multi-word proper name. Fortunately, there is good news too: ambiguous things are usually unambiguously introduced at least once in the text unless they are part of common knowledge presupposed to be known by the readers. This is an observation which can be applied to a broader class of tasks. For example, people are often referred to by their surnames (e.g. "Black") but usually introduced at least once in the text either with their first name ("John Black") or with their title/profession affiliation ("Mr. Black", "President Bush") and it is only when their names are common knowledge that they don't need an introduction (e.g. "Castro", "Gorbachev").

In the case of proper name identification we are not concerned with the semantic class of a name (e.g. whether it is a person name or location) but we simply want to distinguish whether this word in this particular occurrence acts as a proper name or part of a multi-word proper name. If we restrict our scope only to a single sentence, we might find that there is just not enough information to make a confident decision. For instance, *Riders* in the sentence *"Riders said later.."* is equally likely to be a proper noun, a plural proper noun or a plural common noun but if in the same text we find "John Riders" this sharply increases the proper noun interpretation and conversely if we find "many riders" this suggests the plural noun interpretation. Thus our suggestion is to *look at the unambiguous usage of the words in question in the entire document.*

3.1 The Sequence Strategy

Our first strategy for the disambiguation of capitalized words in ambiguous positions is to explore sequences of proper nouns in unambiguous positions. We call it the *Sequence Strategy.* The rationale behind this is that if we detect a phrase of two or more capitalized words and this phrase starts from an unambiguous position we can be reasonably confident that even when the same phrase starts from an unreliable position all its words still have to be grouped together and hence are proper nouns. Moreover, this applies not just to the exact replication of such a phrase but to any partial ordering of its words of size two or more preserving their sequence. For instance, if we detect a phrase *Rocket Systems Development Co.* in the middle of a sentence, we can mark words in the sub-phrases *Rocket Systems*, *Rocket Systems Co.*, *Rocket Co.*, *Systems Development*, etc. as proper nouns even if they occur at the beginning of a sentence or in other ambiguous positions. A span of capital-

		Proper Names				*Common Words*				*Total*	
		All Words		*Known Words*		*All Words*		*Known Words*		*All Words*	
		tokens	*types*	*tokens*	*types*	*tokens*	*types*	*tokens*	*types*	*tokens*	*types*
All Ambiguous		826	339	171	68	1,851	326	1,841	316	2,677	665
Disambiguated	+	795	316	140	54	1,568	218	1,563	213	2,363	534
	−	1	1	1	1	8	8	3	3	9	9
Sequence Strategy	+	62	25	32	11	0	0	0	0	62	25
	−	0	0	0	0	0	0	0	0	0	0
Single Word	+	510	192	108	43	1,270	148	1,265	143	1,780	340
Assignment	−	1	1	1	1	3	3	3	3	4	4
Stop-List	+	0	0	0	0	298	70	298	70	298	70
Assignment	−	0	0	0	0	0	0	0	0	0	0
Lexicon Lookup	+	223	99	0	0	0	0	0	0	223	99
Assignment	−	0	0	0	0	5	5	0	0	5	5
Left Unassigned		30	22	30	22	275	100	275	100	305	122

Table 2: Disambiguated capitalized word-tokens/types in the ambiguous positions.

ized words can also include lower-cased words of length three or shorter. This allows us to capture phrases like *A & M*, *The Phantom of the Opera.*, etc. We generate partial orders from such phrases in a similar way but insist that every generated sub-phrase should start and end with a capitalized word.

To make the Sequence Strategy robust to potential capitalization errors in the document we also use a set of negative evidence. This set is essentially a set of all lower-cased words of the document with their following words (bigrams). We don't attempt here to build longer sequences and their partial orders because we cannot in general restrict the scope of dependencies in such sequences. The negative evidence is then used together with the positive evidence of the Sequence Strategy and block the proper name assignment when controversy is found. For instance, if in a document the system detects a capitalized phrase *"The President"* in an unambiguous position, then it will be assigned as a proper name even if found in ambiguous positions in the same document. To be more precise the method will assign the word "The" as a proper noun since it should be grouped together with the word "President" into a single proper name. However, if in the same document the system detects an alternative evidence e.g. *"the President"* or *"the president"* - it then blocks such assignment as unsafe.

The Sequence Strategy strategy is extremely useful when dealing with names of organizations since many of them are multi-word phrases composed from common words. And indeed, as is shown in Table 2, the precision of this strategy was 100% and the recall about 7.5%: out of 826 proper names in ambiguous positions, 62 were marked and all of them were marked correctly. If we concentrate only on difficult cases when proper names are at the same time common words of English, the recall of the Sequence Strategy rises to 18.7%: out of 171 common words which acted as proper names 32 were correctly marked. Among such words were "News" from "News Corp.", "Rocket" from "Rocket Systems Co.", "Coast" from "Coast Guard" and "To" from "To B. Super".

3.2 Single Word Assignment

The Sequence Strategy is accurate, but it covers only a part of potential proper names in ambiguous positions and at the same time it does not cover cases when capitalized words do not act as proper names. For this purpose we developed another strategy which also uses information from the entire document. We call this strategy *Single Word Assignment*, and it can be summarized as follows: if we detect a word which in the current document is seen capitalized in an unambiguous position and at the same time it is not used lower-cased, this word in this particular document, even when

used capitalized in ambiguous positions, is very likely to stand for a proper name as well. And conversely, if we detect a word which in the current document is used only lower-cased in unambiguous positions, it is extremely unlikely that this word will act as a proper name in an ambiguous position and thus, such a word can be marked as a common word. The only consideration here should be made for high frequency sentence-initial words which do not normally act as proper names: even if such a word is observed in a document only as a proper name (usually as part of a multi-word proper name), it is still not safe to mark it as a proper name in ambiguous positions. Note, however, that these words can be still marked as proper names (or rather as parts of proper multi-word names) by the Sequence Strategy. To build such list of stop-words we ran the Sequence Strategy and Single Word Assignment on the Brown Corpus (Francis&Kucera, 1982), and reliably collected 100 most frequent sentence-initial words.

Table 2 shows the success of the Single Word Assignment strategy: it marked 511 proper names from which 510 were marked correctly, and it marked 1,273 common words from which 1,270 were marked correctly. The only word which was incorrectly marked as a proper name was the word "Insurance" in "Insurance company ..." because in the same document there was a proper phrase "China-Pacific Insurance Co." and no lower-cased occurrences of the word "insurance" were found. The three words incorrectly marked as common words were: "Defence" in "Defence officials ..", "Trade" in "Trade Representation office .." and "Satellite" in "Satellite Business News". Five out of ten words which were not listed in the lexicon ("Pretax", "Benchmark", "Liftoff", "Downloading" and "Standalone") were correctly marked as common words because they were found to exist lower-cased in the text. In general the error rate of the assignment by this method was 4 out of 1,784 which is less than 0.02%. It is interesting to mention that when we ran Single Word Assignment without the stop-list, it incorrectly marked as proper names only three extra common words ("For", "People" and "MORE").

3.3 Taking Care of the Rest

After Single Word Assignment we applied a simple strategy of marking as common words all unassigned words which were found in the stop-list of the most frequent sentence-initial words. This gave us no errors and covered extra 298 common words. In fact, we could use this strategy before Single Word Assignment, since the words from the stop-list are not marked at that point anyway. Note, however, that the Sequence Strategy still has to be applied prior to the stop-list assignment. Among the words which failed to be assigned by either of our strategies were 243 proper names, but only 30 of them were in fact ambiguous, since they were listed in the lexicon of common words. So at this point we marked as proper names all unassigned words which were not listed in the lexicon of common words. This gave us 223 correct assignments and 5 incorrect ones – the remaining five out of these ten common words which were not listed in the lexicon. So, in total, by the combination of the described methods we achieved a *precision* of $\frac{correctly\_assigned}{all\_assigned} = \frac{2363}{2363+9} = 99.62\%$ and a *recall* of $\frac{all\_assigned}{total\_ambiguous} = \frac{2363+9}{2677} = 88.7\%$.

Now we have to decide what to do with the remaining 305 words which failed to be assigned. Among such words there are 275 common words and 30 proper names, so if we simply mark all these words as common words we will increase our recall to 100% with some decrease in precision – from 99.62% down to 98.54%. Among the unclassified proper names there were a few which could be dealt by a part-of-speech tagger: "Gray, chief...", "Gray said...", "Bill Lattanzi...", "Bill Wade...", "Bill Gates...", "Burns , an..." and "..Golden added". Another four unclassified proper names were capitalized words which followed the "U.S." abbreviation e.g. "U.S. Supreme Court". This is a difficult case even for sentence boundary disambiguation systems ((Mikheev, 1998), (Palmer&Hearst, 1997) and (Reynar&Ratnaparkhi, 1997)) which are built for exactly that purpose, i.e., to decide whether a capitalized word which follows an abbreviation is attached to it or whether there is a sentence boundary between them. The "U.S." abbreviation is one of the most difficult ones because it can be as often seen at the end of a sentence as in the beginning of multi-word proper names. Another nine unclassified proper names were stable phrases like "Foreign Minister", "Prime Minister", "Congressional Republicans", "Holy Grail", etc. mentioned just

once in a document. And, finally, about seven or eight unclassified proper names were difficult to account for at all e.g. "Sate-owned" or "Freeman Zhang". Some of the above mentioned proper names could be resolved if we accumulate multi-word proper names across several documents, i.e., we can use information from one document when we deal with another. This can be seen as an extension to our Sequence Strategy with the only difference that the proper noun sequences have to be taken not only from the current document but from the cache memory and all multi-word proper names identified in a document are to be appended to that cache. When we tried this strategy on our test corpus we were able to correctly assign 14 out of 30 remaining proper names which increased the system's precision on the corpus to 99.13% with 100% recall.

4 Discussion

In this paper we presented an approach to the disambiguation of capitalized common words when they are used in positions where capitalization is expected. Such words can act as proper names or can be just capitalized variants of common words. The main feature of our approach is that it uses a minimum of pre-built resources – we use only a list of common words of English and a list of the most frequent words which appear in the sentence-stating positions. Both of these lists were acquired without any human intervention. To compensate for the lack of pre-acquired knowledge, the system tries to infer disambiguation clues from the entire document itself. This makes our approach domain independent and closely targeted to each document. Initially our method was developed using the training data of the MUC-7 evaluation and tested on the withheld test-set as described in this paper. We then applied it to the Brown Corpus and achieved similar results with degradation of only 0.7% in precision, mostly due to the text zoning errors and unknown words. We deliberately shaped our approach so it does not rely on pre-compiled statistics but rather acts by analogy. This is because the most interesting events are inherently infrequent and, hence, are difficult to collect reliable statistics for, and at the same time pre-compiled statistics would be smoothed across multiple documents rather than targeted to a specific document.

The main strategy of our approach is to scan the entire document for unambiguous usages of words which have to be disambiguated. The fact that the pre-built resources are used only at the latest stages of processing (Stop-List Assignment and Lexicon Lookup Assignment) ensures that the system can handle unknown words and disambiguate even very implausible proper names. For instance, it correctly assigned five out of ten unknown common words. Among the difficult cases resolved by the system were a multi-word proper name "To B. Super" where both "To" and "Super" were correctly identified as proper nouns and a multi-word proper name "The Update" where "The" was correctly identified as part of the magazine name. Both "To" and "The" were listed in the stop-list and therefore were very implausible to classify as proper nouns but nevertheless the system handled them correctly. In its generic configuration the system achieved precision of 99.62% with recall of 88.7% and precision 98.54% with 100% recall. When we enhanced the system with a multi-word proper name cache memory the performance improved to 99.13% precision with 100% recall. This is a statistically significant improvement against the bottom-line performance which fared about 94% precision with 100% recall.

One of the key factors to the success in the proposed method is an accurate zoning of the documents. Since our method relies on the capitalization in unambiguous positions – such positions should be robustly identified. In the general case this is not too difficult but one should take care of titles, quoted speech and list entries – otherwise if treated as ordinary text they can provide false candidates for capitalization. Our method in general is not too sensitive to the capitalization errors: the Sequence Strategy is complimented with the negative evidence. This together with the fact that it is rare when several words appear by mistake more than once makes this strategy robust. The Single Word Assignment strategy uses the stop list which includes the most frequent common words. This screens out many potential errors. One notable difficulty for the Single Word Assignment represent words which denote profession/title affiliations. These words modifying

a person name might require capitalization – *"Sheriff John Smith"*, but in the same document they can appear lower-cased – *"the sheriff"*. When the capitalized variant occurs only as sentence initial our method predicts that it should be decapitalized. This, however, is an extremely difficult case even for human indexers – some writers tend to use certain professions such as Sheriff, Governor, Astronaut, etc., as honorific affiliations and others tend to do otherwise. This is a generally difficult case for Single Word Assignment – when a word is used as a proper name and as a common word in the same document, and especially when one of these usages occurs only in an ambiguous position. For instance, in a document about steel the only occurrence of "Steel Company" happened to start a sentence. This lead to an erroneous assignment of the word "Steel" as common noun. Another example: in a document about "the Acting Judge", the word "acting" in a sentence "Acting on behalf.." was wrongly classified as a proper name.

The described approach is very easy to implement and it does not require training or installation of other software. The system can be used as it is and, by implementing the cache memory of multi-word proper names, it can be targeted to a specific domain. The system can also be used as a pre-processor to a part-of-speech tagger or a sentence boundary disambiguation program which can try to apply more sophisticated methods to unresolved capitalized words . In fact, as a by-product of its performance, our system disambiguated about 17% (9 out of 60) of ambiguous sentence boundaries when an abbreviation was followed by a capitalized word.

Apart from collecting an extensive cache of multi-word proper names, another useful strategy which we are going to test in the future is to collect a list of common words which, at the beginning of a sentence, act most frequently as proper names and to use such a list in a similar fashion to the list of stop-words. Such a list can be collected completely automatically but this requires a corpus or corpora much larger than the Brown Corpus because the relevant sentences are rather infrequent. We are also planning to investigate the sensitivity of our method to the document size in more detail.

References

Brill E. 1995 "Transformation-based error-driven learning and natural language parsing: a case study in part-of-speech tagging" In *Computational Linguistics* **21** (4), pp. 543-565

N. Chinchor 1998 Overview of MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference held in Fairfax, VA, April 29-May 1, 1998.* `www.muc.saic.com/muc_7_proceedings/overwiew.html`

K. Church 1995 "One Term Or Two?" In *Proceedings of the 18th Annual Internationals ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95)*, Seattle

K. Church 1988 A Stochastic parts program and noun-phrase parser for unrestricted text. In *Proceedings of the Second ACL Conference on Applied Natural Language Processing (ANLP'88)*, Austin, Texas

W. Francis and H. Kucera 1982 Frequency Analysis of English Usage. Boston MA: Houghton Mifflin.

D. D. Palmer and M. A. Hearst 1997. Adaptive Multilingual Sentence Boundary Disambiguation. In *Computational Linguistics*, **23** (2), pp. 241-269

I. Mani and T.R. MacMillan 1995 Identifying Unknown Proper Names in Newswire Text In B. Boguraev and J. Pustejovsky, eds., *Corpus Processing for Lexical Acquisition*, MIT Press.

M. Marcus, M.A. Marcinkiewicz, and B. Santorini 1993. Building a Large Annotated Corpus of English: The Penn Treebank. In *Computational Linguistics*, vol 19(2), ACL.

A. Mikheev. 1998 "Feature Lattices for Maximum Entropy Modelling" In *Proceedings of the 36th Conference of the Association for Computational Linguistics (ACL/COLING'98)*, pp 848–854. Montreal, Quebec.

A. Mikheev. 1997 "Automatic Rule Induction for Unknown Word Guessing." In *Computational Linguistics* **23** (3), pp. 405-423

A. Mikheev. 1997 "LT POS – the LTG part of speech tagger." Language Technology Group, University of Edinburgh. `www.ltg.ed.ac.uk/software/pos`

A. Mikheev, C. Grover and M. Moens 1998 Description of the LTG system used for MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference held in Fairfax, VA, April 29-May 1, 1998.* `www.muc.saic.com/muc_7_proceedings/ltg-muc7.ps`

J. C. Reynar and A. Ratnaparkhi 1997. A Maximum Entropy Approach to Identifying Sentence Boundaries. In *Proceedings of the Fifth ACL Conference on Applied Natural Language Processing (ANLP'97)*, Washington D.C., ACL.