Exploitation of Named Entities in Automatic Text Summarization for Swedish

Martin Hassel
NADA-KTH
xmartin@nada.kth.se

Abstract

Named Entities are often seen as important cues to the topic of a text. They are among the most information dense tokens of the text and largely define the domain of the text. Therefore, Named Entity Recognition should greatly enhance the identification of important text segments when used by an (extraction based) automatic text summarizer. We have compared Gold Standard summaries produced by majority votes over a number of manually created extracts with extracts created with our extraction based summarization system, SweSum. Furthermore we have taken an in-depth look at how over-weighting of Named Entities affects the resulting summary and come to the conclusion that weighting of Named Entities should be carefully considered when used in a naïve fashion.

Background

The technique of automatic text summarization has been developed for many years (Luhn 1959, Edmundson 1969 and Salton 1989). One way to do text summarization is by text extraction, which means to extract pieces of an original text on a statistical basis or with heuristic methods and put them together to a new shorter text with as much information as possible preserved (Mani & Maybury 1999).

One important task in text extraction is topic identification. There are many methods to perform topic identification (see Lin & Hovy 1997). One is word counting at concept level that is more advanced than just simple word counting; another is identification of cue phrases to find the topic.

To improve our automatic text summarizer and to a larger extent capture the topic of the text we tried to use Named Entity Recognition. Named Entity recognition is the task of finding and classifying proper nouns in running text. Proper nouns, such as names of persons and places, are often central in news reports. Therefore we have integrated a Named Entity tagger with our existing summarizer, SweSum, in order to study its effect on the resulting summaries.

Introducing SweSum

The domain of SweSum (Dalianis 2000) is Swedish newspaper text. SweSum utilizes several different topic identification schemes. For example the bold tag is often used to emphasize contents of the text. Headings are also given a higher weight.

In news paper text the most relevant information is always presented at the top. In some cases the articles are even written to be cuttable from from the bottom. Because of this we use Position Score (Lin & Hovy 1997): sentences in the beginning of the text are given higher scores than later ones.

Sentences that contain keywords are scored high. A keyword is an open class word with a high Term Frequency (*tf*). Sentences containing numerical data are also considered carrying important information.

All the above parameters are put in a naïve combination function with modifiable weights to obtain the total score of each sentence.

Working hypothesis

Named Entities are often seen as important cues to the topic of a text. They are among the most information dense tokens of the text and largely define the domain of the text. Therefore, Named Entity Recognition should greatly enhance the identification of important text segments when used by an (extraction based) automatic text summarizer.

Enter SweNam

For Named Entity recognition and classifying SweNam (Dalianis & Åström 2001) is used. SweNam acts as a preprocessor for SweSum and tags all found Named Entities with one of the four possible categories – names of persons (given name and/or surname), locations (geographical as well as geopolitical), companies (names of companies, brands, products, organizations, etc) and time stamps (dates, weekdays, months, etc).

The Named Entities found by SweNam are quite reliable, as it has shown a precision of 92 percent (Dalianis & Åström 2001). However, the recall is as low as 46 percent, so far from all Named Entities are considered during the summarization phase.

All found entities are given an equal weight and entered, together with the parameters described above, into the combination function in weighting module in the summarizer, SweSum.

Creating a Gold Standard

For the evaluation we collected two sets of texts, each set consisting of 10 news texts. The first set (Group 1) consisted of ten news articles randomly chosen from Svenska Dagbladets web edition (http://www.svd.se/) over a couple of days. These where summarized using SweSum both with and without the use of Named Entity Recognition.

In order to evaluate and compare the two subsets of generated extracts from Group 1 we devised a system to collect manual extracts for the news articles from Group 1. Human test subjects where presented the news articles one at a time in random order in the form of one sentence per line. In front of each line was a checkbox with which the informant could select that particular sentence for extraction. The informant could then choose to generate an extract based on the selected sentences. This extract was then presented to the informant who had to approve the extract before it was entered into a database. Submitted extracts were allowed to vary between 5% and 60% of the original text length.
The result was that 11 informants submitted a total of 96 extracts for the ten texts of Group 1. Each news text received between 8 and 11 manual extracts and the mean length of submitted extracts was 37%.

There was, as expected, not very much agreement between the informants on which sentences to select for the extract. The level of agreement among the informants was calculated with a simple precision function. This is done per text and then a mean was calculated over all ten texts.

$$\frac{Vc}{Ns * Nx} * 100$$

In the function above *Vc* is the number of votes that are represented in the generated extract, *Ns* is the number of sentences represented in the same extract and *Nx* is the number of man-made extracts made for the original text the votes and sentences account for. This means that when all informants choose not only the same number of sentences but also exactly the same set of sentences the function will result in a precision, or agreement, of 100%.

We where prepared for a low agreement among the human extractors as to which sentences are good summary sentences as previous studies have shown this (for an overview see Mani 2001). When taking all selected extraction units into account for each text there was only a mean agreement of 39.6%. This is however not so bad as it can seem at first glance. When generating a "gold standard" extract by presenting the most selected sentences up to a summary length of the mean length of all man-made extracts for a given text the precision, or the agreement level, rose to 68.9%. Very few of the sentences chosen for the gold standard where selected by as few as one third or less of the informants. Of course, even fewer sentences where selected by all informants. In fact, not even all informants could agree upon extracting the title or not when one was present.

Evaluation

The extract summaries generated with SweSum where then manually compared on sentence level with the gold standard extracts generated by majority vote. We found that with Named Entity Recognition the summaries generated by SweSum and the gold standard summaries only had 33.9% of their sentences in common (table 1). On the other hand, without Named Entity Recognition the summaries generated with SweSum shared as many as 57.2% of the sentences with the gold standard.

	With NER	Without NER
Shared sentences	33.9%	57.2%

Table 1: Gold standard compared to SweSum generated extracts

Of course this does not say much about how good the summaries were, only how well the different runs with SweSum corresponded to what our informants wanted to see in the summaries. That is, the figures represent how well SweSum mimics human selection with and without the use of Named Entity Recognition.

Reference errors

The difference in readability and coherence of the two types of SweSum generated summaries was quite interesting. When scrutinizing the extracts we decided to look at a typical problem with extraction-based summarization – reference errors due to removed antecedents. This

error was divided into two severity levels, anaphors that refer to the wrong antecedent and anaphors that does not have any antecedent at all to point to.

In the subset of extracts generated using Named Entity Recognition there where a total of three reference errors (pronouns etc.) and 13 cases of completely lost context over the ten extract summaries (table 2). In the summaries generated not using Named Entity Recognition there were six reference errors and only two cases of completely lost context over the ten summaries.

	With NER	Without NER
Reference errors	3 errors	6 errors
Completely lost context	13 cases	2 cases

Table 2: Referential errors in Group 1 extracts

The extracts generated using Named Entity Recognition clearly showed a lot more coherency problems and loss of context.

To verify the above observations and to see how much NE affected the summarization result we collected a second set of texts (Group 2) and generated new summaries. The second set consisted of 10 news texts randomly chosen from KTH News Corpus (Hassel 2001). These were summarized with a high, low and no weight on Named Entities in SweSum. As shown in table 3 the observations for the Group 1 summaries were very much verified in Group 2. In this new set of extract summaries those generated using Named Entity Recognition showcased a total of 10 respectively 12 reference errors while the set of summaries generated not using Named Entity Recognition only contained 4 errors over the ten summaries.

	High weight on NE	Low weight on NE	No weight on NE
Reference errors	3 errors	3 errors	2 errors
Completely lost context	7 cases	9 cases	2 cases

Table 3: Referential errors in Group 2 extracts

Surprisingly enough the gold standard showed no reference error at all.

Loss of background information

Our conclusion is that weighting of Named Entities tend to prioritize singular sentences high in information centered on the categories used. The result is that it tends to prioritize elaborative sentences over introductory and thus sometimes is responsible for serious losses of sentences giving background information. Our guess is that elaborative sentences have more Named Entities per sentence than introductory due to the fact that introductory sentences focus on something newly introduced in the text. However we have no statistics to substantiate this claim. This often lessens the coherency of the summary (example 1). One solution to this would of course be to extract the paragraph with the highest-ranking sentences (Fuentes & Rodríguez, 2002); another is to let sentence position highly outweigh Named Entities (Nobata et al, 2002).

- Hennes tillstånd är livshotande, säger jourhavande åklagare **Åke Hansson**.
Lisa **Eriksson** var knapphändig i sina uppgifter på tisdagen.
Sjukvården i **Sundsvall** räckte inte till för att rädda flickan.
Enligt läkare i **Uppsala** var hennes tillstånd i går fortfarande livshotande.
2001 anmäldes nära 7 000 fall av barnmisshandel i **Sverig**e. På **Astrid** Lindgrens barnsjukhus i **Solna** upptäcks i dag ungefär ett spädbarn i månaden som är offer för den form av barnmisshandel som kallas Shaken baby-syndrome.
Petter Ovander

Example 1 – Summarized with Named Entities

One way of bouting the problem of loss of background information is of course to raise the size of the extraction unit. If we raise the extraction unit to encompass for example paragraphs instead of sentences the system would identify and extract only the most important paragraph(s) as in Fuentes & Rodríguez (2002). This would lessen the risk of loosing background information at least on paragraph level as well as almost completely eliminate the risk of loss of antecedent for extracted pronouns. On longer texts loss of background information and coherency problem can still of course arise on chapter or text level.

Another way to try to benefit from the use of Named Entity Recognition in Automatic Text Summarization without risking the loss of background information is of course to use a very low weight for NE relative to other weights used (for example keyword frequency and sentence position) and hope that it fine-tunes the summary rather than letting it have a large negative impact on it. This is supported by experiments by Nobata, Sekine, Isahara & Grishman (2002) where they trained an automatic summarization system on English {extract,text} tuples and noted that the weight given by the training system to the Named Entity Recognition module was significantly lower than for the other modules.

Condensed redundancy

When no weighting of Named Entities is carried out clusters of interrelated sentences tend to get extracted because of the large amount of common words. This gives high cohesion throughout the summary but sometimes leads problems with condensed redundancy. For example:

6 veckors baby svårt misshandlad
Pappan misstänkt för misshandeln
En sex veckor gammal bebis kom sent i lördags kväll svårt misshandlad in på akuten i Sundsvall. Flickan har mycket svåra skall- och lungskador. - Hennes tillstånd är livshotande, säger jourhavande åklagare Åke Hansson. Barnets pappa har anhållits som misstänkt för misshandeln på den sex veckor gamla flickan.
Sex veckor gammal
Flickan - som enligt uppgift till Aftonbladet är sex veckor gammal - kom in till akuten Sundsvalls sjukhus vid 22-tiden i lördags kväll. Hennes skador var livshotande.
Petter Ovander

Example 2 – Summarized without Named Entities

We can clearly see how redundancy in the original text “sex veckor gammal” (“six weeks old”) is not only preserved but rather emphasized in the summary. This is because the term frequency (*tf*), the frequency of the keywords, heavy influences the selection.

Over-explicitness

When summarizing with weighting of Named Entities the resulting summaries sometimes seem very repetitive (Example 3) but are in fact generally less redundant than the ones created without weighting of Named Entities.

> Pojkarna skrek att de ville ha pengar och beordrade **Pierre** att gå till kassan.
> **Pierre** minns inte i detalj vad som sedan hände, mer än att det första yxhugget träffade I ryggen.
> Liggande på marken fick **Pierre** ta emot tre yxhugg i huvudet.
> **Pierre** lyckades slita yxan ur händerna på 28-åringen.
> **Pierre** hade svårt att läsa och fick börja om från början igen.
> I dag har **Pierre** lämnat händelserna 1990 bakom sig.
> Psykiskt har **Pierre** klarat sig bra.
>
> **Example 3 – Summarized with Named Entities**

In this case the male name Pierre is repeated over and over again. With the proper noun repeated in every sentence the text appears overly explicit and staccato like. There is no natural flow and the text feels strained and affected. A solution to this would be to generate pronouns in short sequences and keeping only for example every third occurrence of a name in an unbroken name-dropping sequence.

Conclusions

Named Entities, as well as high frequent keywords, clearly carry clues to the topic of a text. Named Entities tend to identify informative extraction segments without emphasizing redundancy by preferring similar segments. A major problem we identified in our experiments is that the Named Entity module tends to prioritize elaborative sentences over introductory and thus sometimes is responsible for serious losses of sentences giving background information. Because of this one of the main difficulties using Named Entities in the weighting scheme would be, as with any lexical or discourse parameter, how to weight it relatively the other parameters. When centering the summary on a specific Named Entity there also arises the need for pronoun generation to avoid staccato like summaries due to over-explicitness.

When producing informative summaries for immediate consumption, for example in a news surveillance or business intelligence system, the background may often be more or less well known. In this case the most important parts of the text is what is new and which participants play a role in the scenario. Here Named Entity Recognition can be helpful in highlighting the different participants and their respective role in the text. Other suggested and applied methods of solving the coherence problem are, as we have seen, to raise the extraction unit to the level of paragraphs or to use a very low, almost insignificant, weight on Named Entities.

Demonstrators

The two different versions of SweSum as well as the small corpus of Swedish news texts and man-made extracts are available on the web if anyone desires to reproduce or do further experiments. The corpus comes with comes with the gold standard extracts generated by majority vote as well as three computer generated baselines. These are available on the following addresses:

SweSum (standard version) – http://swesum.nada.kth.se/index-eng.html
SweSum (NE version) – http://www.nada.kth.se/~xmartin/swesum_lab/index-eng.html
KTH extract corpus - http://www.nada.kth.se/iplab/hlt/kthxc/showsumstats.php

SweNam is also available online for testing purposes:
SweNam – http://www.nada.kth.se/~xmartin/swene/index-eng.html

References

Dalianis H., 2000. *SweSum - A Text Summarizer for Swedish*. Technical report TRITA-NA-P0015, IPLab-174, NADA, KTH.

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

Edmundson H.P., 1969. *New Methods in Automatic Extraction*. Journal of the ACM 16(2) pp 264-285.

Fuentes M. & Rodríguez H., 2002. *Using cohesive properties of text for Automatic Summarization*. JOTRI2002 - Workshop on Processing and Information Retrieval.

Hassel M. 2001. *Internet as Corpus - Automatic Construction of a Swedish News Corpus.* In the Proceedings of NODALIDA '01 - 13th Nordic Conference on Computational Linguistics, May 21-22, 2001, Uppsala, Sweden.

Lin C-Y and Hovy E., 1997. *Identify Topics by Position*. Proceedings of the 5th Conference on Applied Natural Language Processing.

Luhn H.P., 1959. *The Automatic Creation of Literature Abstracts*. IBM Journal of Research and Development pp 159-165.

Mani I., 2001. *Summarization Evaluation: An Overview*. Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization.

Mani I. and Maybury M. (Eds.), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, MA, 1999.

Nobata C., Sekine S., Isahara H. and Grishman R., 2002. *Summarization System Integrated with Named Entity Tagging and IE pattern Discovery*. Proceedings of Third International Conference on Language Resources and Evaluation (LREC 2002); Las Palmas, Canary Islands, Spain.

Salton G., 1989. *Automatic Text Processing: The Transformation Analysis and Retrieval of Information by Computer*. Addison Wesley Publishing Company.