

# Exploitation of Named Entities in Automatic Text Summarization for Swedish

Martin Hassel  
NADA-KTH  
Royal Institute of Technology  
100 44 Stockholm, Sweden  
ph: +46 8 790 66 34  
email: xmartin@nada.kth.se

## 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.

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. 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.

## 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.

## Evaluation

Ten texts were randomly chosen from the web edition of the daily newspaper Aftonbladet ([www.aftonbladet.se](http://www.aftonbladet.se)) and were summarized three times each. Once with no weighting of Named Entities, once with a high weight and once with a relatively low weight.

When no weighting of Named Entities is carried out clusters of interrelated sentences tend to get extracted, which gives high cohesion throughout the summary. 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 1 – Summarized without Named Entities

We can clearly see how redundancy in the original text (“sex veckor gammal”) is not only preserved but rather emphasized in the summary. This is because the term frequency ( $tf$ ) heavily influences the selection.

Weighting of Named Entities instead tend to prioritize singular sentences high in information centered on the categories used. As seen in example 2 below, this often lessens the coherency of the summary. 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 **Sverige**. På **Astrid** Lindgrens barnsjukhus i **Solna** upptäckts i dag ungefär ett spädbarn i månaden som är offer för den form av barnmisshandel som kallas Shaken baby-syndrome.

**Petter Ovander**

### Example 2 – Summarized with Named Entities

These 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. A solution to this would be to generate pronouns in short sequences.

## **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. One of the main difficulties using it 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.

## **References**

- H. Dalianis, 2000. *SweSum - A Text Summarizer for Swedish*. Technical report TRITA-NA-P0015, IPLab-174, NADA, KTH.
- H. Dalianis and E. Åström, 2001. *SweNam - A Swedish Named Entity recognizer. Its construction, training and evaluation*. Technical report, TRITA-NA-P0113, IPLab-189, NADA, KTH.
- H.P. Edmundson, 1969. *New Methods in Automatic Extraction*. Journal of the ACM 16(2) pp 264-285.
- M. Fuentes, H. Rodríguez, 2002. *Using cohesive properties of text for Automatic Summarization*. JOTRI2002 - Workshop on Processing and Information Retrieval.
- C-Y Lin and E. Hovy, 1997. *Identify Topics by Position*. Proceedings of the 5th Conference on Applied Natural Language Processing.
- H.P. Luhn, 1959. *The Automatic Creation of Literature Abstracts*. IBM Journal of Research and Development pp 159-165.
- I. Mani and M. Maybury (Eds.), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, MA, 1999.
- C. Nobata, S. Sekine, H. Isahara and R. Grishman, 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.
- G. Salton, 1989. *Automatic Text Processing: The Transformation Analysis and Retrieval of Information by Computer*. Addison Wesley Publishing Company.