

SweSum – Utilizing Named Entity Recognition in Text Summarization

Martin Hassel

NADA-KTH

Royal Institute of Technology

100 44 Stockholm

ph: +46-8-790 66 34

mail: xmartin@nada.kth.se

Overview of Talk

- Background
- Technique
- Brief study - Observations
- Future improvements
- (Discussion)

Summarization Approaches

- Extraction vs. Abstraction
- Generic vs. Query based (slanted summaries)
- Indicative vs. Informative
- Restricted vs. Unrestricted domain
- Background information vs. New information (TDT)
- Single-document vs. Multiple-document
- Monolingual vs. Multilingual
- Textual vs. Multimedia

Text Summarization

- Extraction is much easier than abstraction
 - Can be extended with rewriting rules: Jing & McKeown [NAACL'00], Jing [ANLP'00]
- Text summarization (extraction) uses statistic, linguistic and heuristic methods

How Is It Done?

- A text is divided into sentences
- Each sentence is assigned a rank
 - Sentence positions (news/reports)
 - Title words
 - Bold text, numerical values, [citation]
 - Named Entities
 - Keyword frequency (nouns, adverbs, adjectives)
 - Use morphological information – lemma

- Words present in many sentences are central
- User adaptation (Query based)
 - Use user keywords - Obtain slanted summaries
- Combination function of all rankings with different weights gives the rank of each sentence
- Generate all high ranking sentences
- Voilà the summary!

SweSum

- The first text summarizer for Swedish
- Summarizes Swedish news paper text in HTML/text format on the WWW.
- Uses a Swedish keyword lexicon that contains 40 000 words and their possible 700 000 inflections.
- During the text summarization 5-10 keywords are produced which describe or categorize the text
 - Keywords - A miniature summary.

Keyword Lexicon

- Keywords in news domain
- Also called "open class word lexicon"
- Keywords can be nouns, adjectives or adverbs

The Keyword Lexicon

700 000 words

40 000 words

Inflected version

Lemma

statsminister

statsminister

statsministern

statsminister

statsministerns

statsminister

statsministrarna

statsminister

statsministrarnas

statsminister

...

...

regeringen

regeringen

regeringens

regeringen

regeringarna

regeringen

regeringarnas

regeringen

...

...

Named Entity Recognition in SweSum

- SweNam - 92 percent precision and 46 percent recall
- Preprocesses the text by tagging found entities with category
- Categories used by default by SweSum:
 - Names (given name and/or surname)
 - Locations (names of places, cities, countries, etc)
 - Companies: (names of companies, brands, products, organizations, etc)[Timestamps are available in SweNam but not used by default by SweSum]
- Found entities are given an equal weight
 - Weighting could be modified by frequency
 - Or chains could be built and weighted

Different Types of Summaries

- Indicative
 - Is this document what I'm looking for?
 - Guide the user to the “right” document
- Informative
 - Making the document “unnecessary”
 - Answer the user's question(s)

Brief Evaluation

- 10 texts of varying length
- News domain
- 3 summaries on each text
 - Named Entities with high weight in ranking
 - Named Entities with low weight in ranking
 - Without using Named Entities
- Comparative study

Summarizing Without Using NE

- Clusters of interrelated sentences tend to get high ranks
- High cohesion throughout the summary

Useful for indicative summaries

Summarizing Using NE

- Tends to prioritize singular sentences
- Lessens the fluency of the summary
- List of facts
 - Can appear to be repetitive but tends to be less redundant

Useful for informative summaries

ORIGINAL TEXT

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.

På onsdagsmorgonen avgör kammaråklagare Lisa Eriksson i Sundsvall om han ska begäras häktad.

Lisa Eriksson var knapphändig i sina uppgifter på tisdagen.

- Utredningen är inne i ett känsligt läge och förhör pågår. Av utredningstekniska skäl kan jag inte berätta mer, säger hon.

Enligt polisens uppgifter var båda föräldrarna med när babyn kom med ambulans till Sundsvalls sjukhus.

-Dagen efter greps pappan av polisen och anhölls senare på sannolika skäl misstänkt för grov misshandel, förklarar Börje Öhman, chef för länskriminalen i Sundsvall.

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. Hon hade bland annat svåra skallskador och svåra skador på lungorna.

Sjukvården i Sundsvall räckte inte till för att rädda flickan. Hon fördes därför med ambulanshelikopter till Akademiska sjukhuset i Uppsala där hon under gårdagen fick specialistvård.

Enligt läkare i Uppsala var hennes tillstånd i går fortfarande livshotande.

Flera fall av våld

Den lilla flickans skador har enligt läkarna orsakats av flera fall av våld i hemmet.

Akademiska sjukhuset ville i natt inte lämna någon ytterligare information om hennes tillstånd.

2001 anmäldes nära 7 000 fall av barnmisshandel i Sverige. 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.

Skadan uppkommer när barnet utsätts för kraftiga skakningar av en vuxen person - oftast en förälder. De barn som överlever riskerar att bli vårdpaket med grava handikapp.

Petter Ovander

Summary using Named Entities

- 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 upptäcks i dag ungefär ett spädbarn i månaden som är offer för den form av barnmisshandel som kallas Shaken baby-syndrome. Petter Ovander
- Lexikon:** Svenska
Ord före 407
Ord efter 121
Sammanfattningsgrad: 29%
Typ av text: tidningstext
Nyckelord: *vecka skada uppgift sjukhus pappa flicka sundsvall akademisk tillstånd*
Hittade **namn:** 1 Åke Hansson; 1 Astrid; 2 Eriksson; 1 Börje Öhmans; 1 Petter Ovander
Hittade **platser:** 4 Sundsvall; 2 Uppsala; 1 Solna; 1 Sverige;

Summary not using Named Entities

- 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
- **Lexikon:** Svenska
Ord före 336
Ord efter 92
Sammanfattningsgrad: 27%
Typ av text: tidningstext
Nyckelord: *vecka skada uppgift pappa flicka sjukhus tillstånd misshandlad misshandeln läkare*

Summary not using Named Entities

- 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 anhängits 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
- **Lexikon:** Svenska
Ord före 336
Ord efter 92
Sammanfattningsgrad: 27%
Typ av text: tidningstext
Nyckelord: *vecka skada uppgift pappa flicka sjukhus tillstånd misshandlad misshandeln läkare*

Solutions?

- Use NE mainly for informative summaries
- Identify and extract only the most important paragraph(s): Fuentes & Rodríguez [JOTRI2002]
- Use rhetorical analysis as well: Alonso Alemany & Fuentes [ESSLLI2002]

Future Improvements of SweSum

- ScandSum
 - Tagging instead of static lexicons
 - Clause level summarization
 - More reliable named entity recognition
 - Improved pronominal resolution
- Lexical chains using SIMPLE and/or EuroWordNet
- Lexical cohesion
- More languages (Norwegian, Farsi)

Demonstrators

- SweSum – Standard version
<http://www.nada.kth.se/~xmartin/swesum/index-eng.html>
- SweSum – Experimental NE version
http://www.nada.kth.se/~xmartin/swesum_lab/index-eng.html
- SweNam
<http://www.nada.kth.se/~xmartin/swene/index-eng.html>