Representing discourse for automatic text summarization via shallow NLP techniques

Laura Alonso i Alemany

Departament de Lingüística General
Universitat de Barcelona
text summarization is...

_A summary is a reductive transformation of a source text into a summary text by extraction or generation_ (Sparck-Jones 2001)
motivation

AS is a challenge for NLP...

• requires a model of human comprehension and production of language
• seems to involve deep processing of language: understanding, generation

... but it also provides a framework to test claims about

• how texts are organized
• the way humans obtain information from texts

discourse for summarization via shallow NLP
1. the problem: automatic text summarization
2. getting to discourse via shallow NLP
3. adequacy of the theoretical model
4. conclusions and future work
contents

1. **the problem: automatic text summarization**
2. getting to discourse via shallow NLP
3. adequacy of the theoretical model
4. conclusions and future work
the problem: automatic text summarization (AS)

no significant improvement in AS since the beginnings of the field (1959)

the performance of baselines is very close to intelligent systems

because summarizing...

... is difficult, requires deep understanding of the texts

... is intrinsically ill-defined

results cannot be properly evaluated,
progress cannot be properly assessed

discourse for summarization via shallow NLP
what works in AS?

very simple techniques:

- extracting **relevant** sentences from the original text
- indicators of relevance:
  - most frequent units (words, sentences), most related units
  - cue phrases: positive (*in sum*), negative (*for example*)
  - position in the document: title, beginning of document, paragraph
  - presence of definite information: Named Entities, dates...

 niche improvement in the quality of summaries

discourse for summarization via shallow NLP
why discourse?

useful for AS
why discourse?

useful for AS

In this case, and thanks to the excellent work of the anthropologist Silvia Ventosa, author of “Work and life of Barcelona’s seamstresses”, this urban legend was found to be a copy of a story that shook the French town of Orleans in 1969.
why discourse?

useful for AS

[1 En este caso,]
[2 [y] [gracias al] excelente trabajo de la antropóloga Silvia Ventosa,]
[5 autora de "Trabajo y vida de las corseteras de Barcelona",]
[6 esta leyenda urbana se comprobó]
[7 que era un calco de una historia]
[8 que conmocionó a la localidad francesa de Orleans]
[9 en 1969].
discourse for summarization via shallow NLP
why discourse?

useful for AS

summary:
esta leyenda urbana se comprobó que era un calco de una historia
this urban legend was found to be a copy of a story

discourse for summarization via shallow NLP
why discourse?

useful for AS

general  not particular for a genre,

insightful  provides a high-level analysis of the structure of texts

robust  can be represented with shallow NLP techniques
1. the problem: automatic text summarization
2. **getting to discourse via shallow NLP**
3. adequacy of the theoretical model
4. conclusions and future work
questions to be answered

• what is an adequate representation of discourse?
• what is the linguistic unit at discourse level?
• which relations can be established between discourse units?
• how do we exploit a representation of discourse for AS?
getting to discourse via shallow NLP

top-down targeted representation

- discourse units segments and operators
- relations between units
- purpose of the analysis assess relevance and coherence for AS

bottom-up an inventory of discursive meanings based on recognizable clues

framework current NLP capabilities for Catalan and Spanish

discourse for summarization via shallow NLP
targeted representation

- a directed acyclic graph: a list of trees
- multidimensional: heterogeneous dimensions of meaning represented separately
an inventory of discursive meanings

**compositional** vs. **atomic**

**multidimensional** vs. **holistic**

**data-driven** vs. determined by a theory

**general** vs. **ad-hoc**
surface clues about discourse organization

- punctuation
- shallow syntactical structures
- discourse markers

very rich in discursive semantics, systematized by semantic maps

discourse for summarization via shallow NLP
an inventory of discursive meanings

structural dimension

  continuation
  elaboration

semantic dimension

  revision
  causality
  equality
  context

discourse for summarization via shallow NLP
# An inventory of discursive meanings

<table>
<thead>
<tr>
<th>feature</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>revision</td>
<td>negates some content from the previous discourse</td>
</tr>
<tr>
<td></td>
<td><em>she is serious</em> <strong>but nice.</strong></td>
</tr>
<tr>
<td>equivalence</td>
<td>establishes an equivalence between two units</td>
</tr>
<tr>
<td></td>
<td><em>we cook lots of vegetables, for example, spinach.</em></td>
</tr>
<tr>
<td>causality</td>
<td>elicits a causal relation between two units</td>
</tr>
<tr>
<td></td>
<td><em>it fell because it was in the wrong place.</em></td>
</tr>
<tr>
<td>context</td>
<td>provides background for a discourse entity</td>
</tr>
<tr>
<td></td>
<td><em>we arrived home when it began raining.</em></td>
</tr>
<tr>
<td>elaboration</td>
<td>continues an already presented topic or intention</td>
</tr>
<tr>
<td></td>
<td><em>you will like my mom, she’s nice, knows stories...</em></td>
</tr>
<tr>
<td>continuation</td>
<td>introduces a new topic or intention</td>
</tr>
<tr>
<td></td>
<td><em>you will like my mom, and you must meet my father.</em></td>
</tr>
</tbody>
</table>

discourse for summarization via shallow NLP
### an inventory of discursive meanings

<table>
<thead>
<tr>
<th>relation</th>
<th>semantic</th>
<th>structural</th>
</tr>
</thead>
<tbody>
<tr>
<td>contrast</td>
<td>revision</td>
<td>continuation</td>
</tr>
<tr>
<td>concession</td>
<td>revision</td>
<td>elaboration</td>
</tr>
<tr>
<td>result-driven cause</td>
<td>cause</td>
<td>continuation</td>
</tr>
<tr>
<td>reason</td>
<td>cause</td>
<td>elaboration</td>
</tr>
<tr>
<td>parallelism</td>
<td>equality</td>
<td>continuation</td>
</tr>
<tr>
<td>exemplification</td>
<td>equality</td>
<td>elaboration</td>
</tr>
<tr>
<td>topic</td>
<td>context</td>
<td>continuation</td>
</tr>
<tr>
<td>background</td>
<td>context</td>
<td>elaboration</td>
</tr>
<tr>
<td>narration</td>
<td>narration</td>
<td>continuation</td>
</tr>
<tr>
<td>explanation</td>
<td>narration</td>
<td>elaboration</td>
</tr>
</tbody>
</table>

**discourse for summarization via shallow NLP**
1. the problem: automatic text summarization
2. getting to discourse via shallow NLP
3. adequacy of the theoretical model
4. conclusions and future work
adequacy of the theoretical model

- to the opinion of human judges
- to automatic procedures
- to text summarization
adequacy of the theoretical model to summarization

combination of lexical chains and discourse analysis

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>Cosine similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical Chains</td>
<td>.73</td>
<td>.74</td>
<td>.81</td>
</tr>
<tr>
<td>+ Discourse Semantic</td>
<td>.74</td>
<td>.76</td>
<td>.82</td>
</tr>
<tr>
<td>+ Semantic + Structural</td>
<td>.79</td>
<td>.80</td>
<td>.84</td>
</tr>
</tbody>
</table>
Adequacy of the theoretical model to summarization

Combination of lexical chains and discourse analysis

Integration in an e-mail summarizer

Discourse structural information is relevant to summarize 80% of e-mails, with a precision of .63 in comparison with the gold standard

- Baselines: first paragraph: .35, first sentence: .65
- Kappa agreement between annotators: .5
1. the problem: automatic text summarization
2. getting to discourse via shallow NLP
3. adequacy of the theoretical model
4. conclusions and future work
contribution

• provide a framework to represent discourse with shallow NLP techniques
  – a formal structure for the representation of discourse
  – a data-driven inventory of basic meanings
  – a relation between surface clues and discourse meanings

• resources for NLP applications
  – annotated corpora
  – a trilingual lexicon of discourse markers
  – algorithms for discourse segmentation with various information
  – a tool for automatic acquisition of discourse markers in raw corpus

discourse for summarization via shallow NLP
future work

• apply the presented methodology to other languages, to
  – provide support or counterevidence to the inventory here
  – develop resources for NLP

• develop a discourse shallow parser
*References