Further Explorations in Sentence Fusion

Daeso

Emiel Krahmer
[joint work with Erwin Marsi]

Twente, March 27, 2008
Plan

1. Introduction: A short history of sentence fusion research

2. Part I: Developing a Parallel Monolingual Treebank

3. Part II: Q-driven vs. Generic sentence fusion
   3a. Experiment 1: Data-collection
   3b. Experiment 2: Evaluation

4. Summary and outlook
Plan

1. Introduction: A short history of sentence fusion research

2. Part I: Developing a Parallel Monolingual Treebank

3. Part II: Q-driven vs. Generic sentence fusion
   3a. Experiment 1: Data-collection
   3b. Experiment 2: Evaluation

4. Summary and outlook
Motivation: Semantic Overlap

- Singer Christina Aguilera has finally revealed why her stomach is so round. [NOS]

- Christina Aguilera, in the American magazine Glamour, has confirmed that she is pregnant. [AD]

- Christina Aguilera has finally asserted what the whole world already knew: she is expectant. [NOVUM]

- Everyone knew, but this time Christina Aguilera is said to have confirmed it for the first time: she is pregnant. [The Agenda]
Why bother?

- Similar information can be expressed ("paraphrased") in many different ways.

- Major **stumbling block** for robust NLP applications such as summarization (Bosma 2008), but also IE, IR or QA.

- **Resources** exist on the word level (e.g., Wordnet), but are mostly lacking for more complex phrases.

- The Stevin **Daeso** (Detecting and Exploiting Semantic Overlap) project [a direct Imogen-IMIX spin-off] intends to fill this gap.
Sentence Fusion

- **Sentence fusion**: given two related sentences, produce a single sentence with the same information (Barzilay et al. 1999, Barzilay & McKeown 2005)

- **Example**:
  - Christina Aguilera has confirmed, in the American magazine Glamour, that she is pregnant.
  - Christina Aguilera has finally asserted what the whole world already knew: she is expectant.

- **Fusion**: Christina Aguilera has confirmed that she is pregnant.

- **Motivation**: Beneficial for multi-document summarization. Less redundancy, more informative summaries (Barzilay & McKeown 2005)
General strategy

- **Analysis**
  - Alignment of dependency structures.

- **Merge**
  - Merging dependency trees.

- **Generation**
  - Use *n*-gram models to rank generated sentences.

S1 \(\rightarrow\) Analysis \(\rightarrow\) Merge \(\rightarrow\) Generation \(\rightarrow\) Fused sentence

S2
First complication

- Daume III & Marcu (2004): “Generic sentence fusion is an ill-defined summarization task.”

- When participants are asked to fuse two consecutive sentences from a document, their results are widely different.

- If human participants don’t agree, evaluating sentence fusion is tricky...
Second complication

- Marsi & Krahmer (2005): There is more than one way to fuse two sentences.

- Reconsider:
  - Christina Aguilera has confirmed, in the American magazine Glamour, that she is pregnant.
  - Christina Aguilera has finally asserted what the whole world already knew: she is expectant.

- Intersection Fusion: Christina Aguilera has confirmed that she is pregnant.

- Union Fusion: Christina Aguilera has finally confirmed, in the American magazine glamour, what the whole world already knew: she is pregnant.

- Which type of fusion is best for a particular application is an open question...
Three Issues

- **Issue #1**: Fusions of different levels (union, intersection, ...) require a better understanding of how different words and phrases in different sentences may be related.
  - **Part I**: Ongoing effort to build a Parallel Monolingual Treebank (for Dutch)

- **Issue #2**: Is Q-based fusion a better defined task? Will people agree more for unions than for intersections? Is the effect of the preceding question the same for both unions and intersections.

- **Issue #3**: And: which would people prefer? Would they prefer short (intersection) or extensive (union) answers? And does it matter whether they were generic of Q-based?
  - **Part II**: Two evaluation experiments which address these questions.
Plan

1. Introduction: A short history of sentence fusion research

2. Part I: Developing a Parallel Monolingual Treebank

3. Part II: Q-driven vs. Generic sentence fusion
   3a. Experiment 1: Data-collection
   3b. Experiment 2: Evaluation

4. Summary and outlook
The Daeso corpus and tools

- Building a 1M word parallel monolingual treebank.

- Basic idea: look for pairs of sentences where there is an independent criterium that there will be some amount of overlap.

- The corpus should contain different text genres and different amounts of overlap.

- 500K manually aligned and corrected [now]; 500K automatic [2009]
Corpus collection

- **Autocue - Subtitling** (NOS, TwNC)

- Parallel translations into Dutch

- Google **Headlines** (mined by Wauter Bosma)

- Different press releases (ANP, Novum) about the same (Dutch) event.

- Potential sets of **answers** (automatically found) to different questions (used Joost QA, Bouma et al. 2007).
## Corpus data

<table>
<thead>
<tr>
<th></th>
<th>Manual</th>
<th>Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autocue-subtitles</td>
<td>125k</td>
<td>192k</td>
</tr>
<tr>
<td>Book translations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Darwin1</td>
<td>25k</td>
<td>154k</td>
</tr>
<tr>
<td>Darwin 2</td>
<td>25k</td>
<td>191k</td>
</tr>
<tr>
<td>Montaigne 1</td>
<td>25k</td>
<td>462k</td>
</tr>
<tr>
<td>Montaigne 2</td>
<td>25k</td>
<td>~500k</td>
</tr>
<tr>
<td>Saint-Exupéry 1</td>
<td>15k</td>
<td>15k</td>
</tr>
<tr>
<td>Saint-Exupéry 2</td>
<td>15k</td>
<td>15k</td>
</tr>
<tr>
<td>News headlines</td>
<td>24k</td>
<td>&gt; 900k</td>
</tr>
<tr>
<td>Press releases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANP</td>
<td>125k</td>
<td>197k</td>
</tr>
<tr>
<td>Novum</td>
<td>125k</td>
<td>136k</td>
</tr>
<tr>
<td>QA system output</td>
<td>1k</td>
<td>1k</td>
</tr>
</tbody>
</table>
Pre-processing and annotation steps

1. XML TEI format (Text Encoding Initiative).

2. Sentence splitting and tokenization with the DCOI tokenizer for Dutch (Reynaert 2007).

3. Dependency parsing with the Alpino parser (van Noord et al.).

4. Alignment at text and sentence level.

5. Alignment of dependency trees.
Sentence alignment

- Standard alignment methods (e.g., Gale and Church 1993) assume alignment is mostly 1-to-1 and that crossing alignments and unaligned sentences are rare.

- These assumptions are often violated.
  - Obviously in comparable texts
  - But also in e.g., translations of Darwin’s Origin of Species

- Developed
  - A new alignment method to boost manual annotation.
  - A new annotation tool to check sentence alignments
Automatic sentence alignment

- Tricky for comparable texts

- As a first approximation: low level, multiple pass, shallow features.

- Experiments with:
  - types vs token;
  - different overlap metrics (MaxSim, Cosine, Jaccard, Dice, Tanimoto, ...);
  - tf-idf weighting (Nelken & Schieber 2006)

- Ongoing...
Hitaext: Tool for text and sentence alignment

First public release (October 2007): http://daeso.uvt.nl/hitaext/
Alignments of words and phrases

- Given two dependency trees for two aligned sentences: align nodes and label the alignment relation.
  - “Christina Aguilera” equals “Christina Aguilera”
  - “pregnant” restates “expectant”
  - “the singer Aguilera” specifies “Aguilera”
  - “Aguilera” generalizes “the singer Aguilera”
  - “Christina Aguilera and Beyoncé” intersects “Beyoncé and Pink”

- Marsi & Krahmer (2005): for first five chapters of “Le Petit Prince”, two annotators reached an F-score of .98 on relations and .95 on labels.

- Recently replicated with independent annotators
Algraeph: Tool for aligning nodes and labeling alignments

First public release (March 2008): http://daeso.uvt.nl/algraeph/
State of affairs

- Work on manual alignment of words and phrases currently ongoing.

- Other work on the corpus is now finished.

- Further activities:
  - Multi-document summarization: currently building a baseline multi-document summarization system for Dutch, to be extended with Daeso tools later on.
  - Sentence fusion: combine two related sentences into a single grammatical sentence. New results on question-driven fusion
Plan

1. Introduction: A short history of sentence fusion research

2. Part I: Developing a Parallel Monolingual Treebank

3. Part II: Q-driven vs. Generic sentence fusion
   3a. Experiment 1: Data-collection
   3b. Experiment 2: Evaluation

4. Summary and outlook
Experiment 1: Data collection

- **Materials:**
  - Used the IMIX QA evaluation set (100 questions).
  - Correct answers were manually retrieved from the IMIX corpus.
  - Selected 25 questions which resulted in multiple answers, which could be union fused [trivial] and intersected.

- **Mixed between-within participants design.** Two between conditions: Intersection and Union. Within each condition, both Generic and Question-based.

- **Participants:** 44 participants (24 men), average age 30.1 years. Randomly assigned to conditions.

- **Method:** web-based script.
Example

Q: What is PTSD?

- A1: Posttraumatic stress disorder (PTSD) is a psychological disorder which is classified as an anxiety disorder in the DSM-IV.

- A2: Posttraumatic stress disorder (abbrev. PTSD) is a psychological disorder caused by a mental trauma (also called psychotrauma) that can develop after exposure to a terrifying event.

So: half of the participants produce intersections of pairs such as these, and half produce unions. All see the pairs first without, than with relevant question.
Results (1)

- Descriptive statistics

<table>
<thead>
<tr>
<th>Fusion Type</th>
<th>Length M (SD)</th>
<th># Ident.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q-based Intersection</td>
<td>8.1 (2.5)*</td>
<td>189*</td>
</tr>
<tr>
<td>Generic Intersection</td>
<td>15.6 (2.9)</td>
<td>73</td>
</tr>
<tr>
<td>Q-based Union</td>
<td>19.2 (4.7)*</td>
<td>134^</td>
</tr>
<tr>
<td>Generic Union</td>
<td>31.2 (7.8)</td>
<td>109</td>
</tr>
</tbody>
</table>

* p < .001, ^ n.s.
## Results (2)

- (Normalized) ROUGE scores

<table>
<thead>
<tr>
<th></th>
<th>Generic Intersection</th>
<th>Q-based Intersection</th>
<th>Generic Union</th>
<th>Q-based Union</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rouge-1</td>
<td>.036</td>
<td>.068</td>
<td>.035</td>
<td>.041</td>
</tr>
<tr>
<td>Rouge-SU4</td>
<td>.014</td>
<td>.038</td>
<td>.018</td>
<td>.020</td>
</tr>
<tr>
<td>Rouge-SU9</td>
<td>.014</td>
<td>.040</td>
<td>.016</td>
<td>.020</td>
</tr>
</tbody>
</table>
In sum

- Q-based fusions are shorter, display less variety in length, yield more identical results, and have higher normalized ROUGE scores.

- So: Q-based fusion is indeed a better defined task.

- But: does it matter?
Experiment 2: Evaluation

Materials:
- Selected 20 questions for which multiple (different) answers were obtained in Experiment I.
- Per questions, 4 representative answers were selected from the data collection, one for each category: Q-based Intersection, Q-based Union, Generic Intersection, Generic Fusion.

Within participants design. For each of the 20 questions, participants have to rank the four answer

Participants: 38 participants (17 men), average age 39.4 years.

Method: simulated medical QA system.
What is PTSD?

- [Generic Intersection] Posttraumatic stress disorder (PTSD) is a psychological disorder.

- [Q-based Intersection] PTSD stands for posttraumatic stress disorder and is a psychological disorder.

- [Generic Union] Posttraumatic stress disorder (PTSD) is a psychological disorder, which is classified as an anxiety disorder in the DSM-IV, caused by a mental trauma (also called psychotrauma) that can develop after exposure to a terrifying event.

- [Q-based Union] PTSD (posttraumatic stress disorder) is a psychological disorder caused by a mental trauma (also called psychotrauma) that can develop after exposure to a terrifying event.
Results

- Average rank

<table>
<thead>
<tr>
<th>Rank</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Q-based Union</td>
<td>1.888*</td>
</tr>
<tr>
<td>2</td>
<td>Q-based Intersection</td>
<td>2.471*</td>
</tr>
<tr>
<td>3</td>
<td>Generic Intersection</td>
<td>2.709*</td>
</tr>
<tr>
<td>3</td>
<td>Generic Union</td>
<td>2.932</td>
</tr>
</tbody>
</table>

* $p < .001$
In sum

- Q-based answer fusions are systematically preferred over generic ones.

- More complete answers are preferred over shorter ones (consistent with Bosma 2008).
Plan

1. Introduction: A short history of sentence fusion research

2. Part I: Developing a Parallel Monolingual Treebank

3. Part II: Q-driven vs. Generic sentence fusion
   3a. Experiment 1: Data-collection
   3b. Experiment 2: Evaluation

4. Summary and outlook
Three issues

- **#1**: More knowledge/data is required about semantic overlap between sentences.

- Daeso corpus does contain this data, and work on the corpus is steadily progressing.

- **Next steps:**
  - Improve tools for detecting and exploiting semantic overlap
  - Apply in context of various applications (multi-document summarization, QA, IE)
Three issues (cont.)

- **#2: Is Q-based fusion a better defined task?**
  Yes. Q-based fusions are shorter, less varied, yield more identical solutions and have higher (normalized) Rouge scores than their generic counterparts.

- **#3: Which type of fusions of users prefer in a QA context?**
  Q-based Union >> Q-based Intersections >> Generic Fusions

- **Next steps:**
  - Follow-up experiments looking at the influence of question wording and at different text genres
  - Working on extended fusions algorithm, based on Marsi & Krahmer (2005)
About the Daeso Stevin project

- **People involved:** Hanneke Schoormans, Nienke Eckhardt, Paul van Pelt, Jurry de Vos, Iris Hendrickx, Walter Daelemans, Jakub Zavrel, Maarten de Rijke, Erwin Marsi, Emiel Krahmer

- **More info:**

  http://daeso.uvt.nl/