Part I:
The DAESO Project - Detecting and Exploiting Semantic Overlap

Erwin Marsi & Emiel Krahmer
Dept. of Communication and Information Sciences
Tilburg University
The Netherlands
Semantic Overlap

• Steve Irwin, the TV host known as the "Crocodile Hunter," has died after being stung by a stingray off Australia's north coast. [www.cnn.com]

• Steve Irwin, the daredevil wildlife documentarian, is killed in a stingray attack while filming on the Great Barrier Reef. [www.time.com]
The DAESO Project

• DAESO: Detecting And Exploiting Semantic Overlap
• Funded by the Dutch Stevin programme
• Duration: October 2006 until October 2009
• Participants:
  – Emiel Krahmer, Erwin Marsi - Tilburg University
  – Walter Daelemans, Iris Hendrickx - Antwerp University
  – Maarten de Rijke, Erik TKS - University of Amsterrdam
  – Jakub Zavrel - Textkernel
• http://daeso.uvt.nl/
1. Development of a corpus of parallel/comparable Dutch text aligned at the level of sentences, phrases and words.

2. Based on this, development of tools for automatic alignment and detection of semantic overlap.

3. Application of these tools in a number of NLP applications.
**Corpus: Text Material**

<table>
<thead>
<tr>
<th>Targets for manual annotation</th>
<th>#words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book translations</td>
<td>125k</td>
</tr>
<tr>
<td>Autocue-subtitle pairs</td>
<td>125k</td>
</tr>
<tr>
<td>News headlines</td>
<td>24k</td>
</tr>
<tr>
<td>Press releases</td>
<td>225k</td>
</tr>
<tr>
<td>Answers from QA</td>
<td>1k</td>
</tr>
<tr>
<td><strong>Subtotal manually annotated</strong></td>
<td><strong>500k</strong></td>
</tr>
</tbody>
</table>

**Target for automatic annotation:** 500k

**Target final corpus:** 1M
Corpus: Annotation Steps

1. Conversion to electronic format (XML)
2. Tokenization & sentence splitting
3. Alignment at text level
4. Alignment at sentence level
5. Syntactic parsing
6. Alignment of syntactic trees
Software Tools: So Far

• Tools for manual alignment:
  – Hitaext: graphical tool for manually aligning pairs of text with XML markup
  – Algraephe: graphical tool for manually aligning and nodes from a pair of graphs, and labeling the alignment relations

• A number of quick & dirty tools for automatic alignment to facilitate manual annotation
Software Tools: To Do

Use the corpus to develop software for:

• Automatic preprocessing, sentence alignment and tree alignment
• Automatic paraphrase extraction
• Sentence fusion: merging trees and generating a new sentence (Barzilay & McKeown)
Sentence Fusion

**Input sentences:**

1. Steve Irwin, the TV host known as the "Crocodile Hunter", has died after being stung by a stingray off Australia's north coast.
2. Steve Irwin, the daredevil wildlife documentarian, is killed in a stingray attack while filming on the Great Barrier Reef.

**Intersection fusion:**

Steve Irwin is killed in a stingray attack off Australia's north coast.

**Union fusion:**

Steve Irwin, the TV host known as the "Crocodile Hunter" and daredevil wildlife documentarian, has died after being stung by a stingray while filming on the Great Barrier Reef.
Sentence fusion

1st sentence

Analysis

Merge

Generation

2nd sentence

analysing and aligning sentences

merging sentences

generating new sentence

Fused sentence
Applications

Exploit semantic overlap in:
• Multi-document summarization
• Question-Answering
• Information Extraction
Multi-document Summarization

Given a set of similar documents

1. Rank sentences on informativeness
2. Align similar sentences
3. Extract sentences to create summary (extract)
4. Revise summary text (abstract)
Summary of the DAESO project

The DAESO (Detecting And Exploiting Semantic Overlap) project concerns:

1. Development of a 1M words corpus of parallel/comparable Dutch text aligned at the level of sentences, phrases and words - plus tools and guidelines for manual annotation

2. Based on this, development of tools for automatic alignment, detection of semantic overlap, paraphrase extraction and sentence fusion

3. Application of these tools in a number of NLP applications, including multi-doc sum, QA and IE.
Part II: Shallow approaches to sentence alignment in comparable text

Erwin Marsi & Emiel Krahmer
Dept. of Communication and Information Sciences
Tilburg University
The Netherlands
**Plan**

- Task definition, example
- Motivation
- Experimental setting
  - Data set
  - Alignment procedure
  - Evaluation method
- Experimental results
- Conclusion, discussion, future work
Sentence Alignment in Comparable Text

Task definition:

given two comparable texts,
align all sentence pairs which share similar information
<table>
<thead>
<tr>
<th>Source Text</th>
<th>Target Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. An attack by a large swarm of jellyfish has wiped out more than 10,000 salmon at a fish farm in Northern Ireland.</td>
<td>1. A 10-square-mile pack of jellyfish wiped out a 100,000-fish salmon farm in Northern Ireland, the Associated Press reported Wednesday.</td>
</tr>
<tr>
<td>2. The manager of the farm says billions of jellyfish turned the sea red, covering an area of 25 square kilometres.</td>
<td>2. The jellyfish covered an area of around 25 sq km and engulfed the Northern Salmon company's cages</td>
</tr>
<tr>
<td>3. He says staff were unable to do anything to prevent the attack, which lasted for several hours.</td>
<td>3. The billions of jellyfish, piled densely in a 35-foot-deep layer, did in the fish through stings and stress, according to John Russell, managing director of Northern Salmon.</td>
</tr>
<tr>
<td>4. Scientists believe jellyfish swarms are increasing because of the depletion of other fish stocks.</td>
<td>4. Scientists pointed to the presence of the jellyfish, rarely seen that far north, as evidence of global warming.</td>
</tr>
</tbody>
</table>
Source Text

1. An attack by a large swarm of jellyfish has wiped out more than 10,000 salmon at a fish farm in Northern Ireland.
2. The manager of the farm says billions of jellyfish turned the sea red, covering an area of 25 square kilometres.
3. He says staff were unable to do anything to prevent the attack, which lasted for several hours.
4. Scientists believe jellyfish swarms are increasing because of the depletion of other fish stocks.

Target Text

1. A 10-square-mile pack of jellyfish wiped out a 100,000-fish salmon farm in Northern Ireland, the Associated Press reported Wednesday.
2. The jellyfish covered an area of around 25 sq km and engulfed the Northern Salmon company's cages.
3. The billions of jellyfish, piled densely in a 35-foot-deep layer, did in the fish through stings and stress, according to John Russell, managing director of Northern Salmon.
4. Scientists pointed to the presence of the jellyfish, rarely seen that far north, as evidence of global warming.
1. An attack by a large swarm of jellyfish has wiped out more than 10,000 salmon at a fish farm in Northern Ireland.

2. The manager of the farm says billions of jellyfish turned the sea red, covering an area of 25 square kilometres.

3. He says staff were unable to do anything to prevent the attack, which lasted for several hours.

4. Scientists believe jellyfish swarms are increasing because of the depletion of other fish stocks.

1. A 10-square-mile pack of jellyfish wiped out a 100,000-fish salmon farm in Northern Ireland, the Associated Press reported Wednesday.

2. The billions of jellyfish, piled densely in a 35-foot-deep layer, did in the fish through stings and stress, according to John Russell, managing director of Northern Salmon.

3. The jellyfish covered an area of around 25 sq km and engulfed the Northern Salmon company's cages.

4. Scientists pointed to the presence of the jellyfish, rarely seen that far north, as evidence of global warming.
**Sentence Alignment in Comparable Text**

### Source Text

1. An attack by a large swarm of jellyfish has wiped out more than 10,000 salmon at a fish farm in Northern Ireland.
2. The manager of the farm says billions of jellyfish turned the sea red, covering an area of 25 square kilometres.
3. He says staff were unable to do anything to prevent the attack, which lasted for several hours.
4. **Scientists believe jellyfish swarms are increasing because of global warming.**

### Target Text

1. A 10-square-mile pack of jellyfish wiped out a 100,000-fish salmon farm in Northern Ireland, the Associated Press reported Wednesday.
2. The billions of jellyfish, piled densely in a 35-foot-deep layer, did in the fish through stings and stress, according to John Russell, managing director of Northern Salmon.
3. The jellyfish covered an area of around 25 sq km and engulfed the Northern Salmon company's cages.
4. **Scientists pointed to the presence of the jellyfish, rarely seen that far north, as evidence of global warming.**
Sentence Alignment in Parallel vs. Comparable Text

**Sentence alignment in parallel text**
- Cleary defined task
- Relatively easy task
- Has established methods (e.g. Church & Gale)
- Domain-independent and knowledge-lean methods
- Large-scale data sets

**Sentence alignment in comparable text**
- Task definition arguably vague
- Relatively hard task
- Methods for *parallel* text fail; requires different approach (Barzilay & Elhadad, Nelken & Shieber)
- No real generic methods yet
- Small data sets
Shallow versus Deep NLP

**Shallow processing**
- Data-driven (empirical)
- Requires training data
- Language-independent
- Domain-independent
- Robust (fault-tolerant, graceful degradation)
- Computationally cheap
- Possibly black box

**Deep processing**
- Knowledge-based (theoretical)
- Requires (computer) linguists
- Language-specific
- Domain-specific
- Brittle
- Computationally expensive
- Hopefully transparent
Shallow sentence alignment?

Our motivation for studying shallow sentence alignment:

1. Shallow approach sets a sharp baseline
2. Shallow processing can quickly filter out the "easy" cases
3. Deep processing may be too expensive for on-the-fly alignment of substantial texts
Data set

- From the DAESO corpus
- Press releases from two Dutch news agencies
- 693 pairs of comparable news texts manually aligned by two human annotators
- Sentence splitting performed automatically
- Sentences manually aligned by two human annotators
- Sentences may be unaligned
- Alignments may be many-to-many
Automatic Alignment procedure

Pseudo-code for automatic alignment procedure:

```python
for (SourceText, TargetText) in AlignedTextPairs:
    for SourceSent in SourceText:
        for TargetSent in TargetText:
            if Sim(SourceSent, TargetSent) > Th:
                Align(SourceSent, TargetSent)
```
Evaluation: Precision, Recall and F-score

$True :=$ set of all manually aligned sentence pairs from the reference corpus

$Pred :=$ set of all aligned sentence pairs as predicted

$$\text{precision} = \frac{|True \cap Pred|}{|Pred|}$$

$$\text{recall} = \frac{|True \cap Pred|}{|True|}$$

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$
Types and Tokens

- Finding similar sentences amount to finding similar parts
- Naive unit is **word**
  - Exact definition turns out to be hard
- Practical unit is **token**
  - Basic definition: everything separated by white space
  - Usually with punctuation detached from the word
- Chopping a sentence into tokens is called **tokenization**
- Tokens can be regarded as a **bag** (multiset)
- Considering tokens as a **set** gives you **types**

String: Here is a "random" example of a sentence.
Tokens: bag {Here is a " random " example of a sentence .
Types: set {Here is a " random example of sentence}
Similarity Measures: TokenMatch

\[ \text{TokenMatch}(A,B) = \left| \{ t \in A \mid t \in B \} \right| \]

where A and B are bags of tokens

- This is simply the overlap in tokens
- Relation is not symmetric
  - \( \text{TokenMatch}([\text{A}, \text{A}], [\text{A}]) = 2 \)
  - \( \text{TokenMatch}([\text{A}], [\text{A}, \text{A}]) = 1 \)

\[ \text{SymTokenMatch}(A,B) = \frac{\text{TokenMatch}(A,B) + \text{TokenMatch}(B,A)}{2} \]
**Similarity Measures: TypeMatch**

\[ TypeMatch(A, B) = |A \cap B| \]

where A and B are sets of tokens

- This is simply overlap in types
- Relation is symmetric
- In practice often no different from TokenMatch
Results for *TokenMatch* and *TypeMatch*

Observation:
- matching on types works better than matching on tokens
Preprocessing

- Preprocessing to reduce superficial differences:
  1. Strip punctuation
  2. Remove function words
     - defined as the closed-class words, i.e. determiners, pronouns, prepositions, auxiliary verbs, connectives, complementizers, and some
  3. Remove domain-dependent words
     - e.g. names of press agencies
  4. Lower case all tokens

- May introduce errors, but overall effect is probably positive
Observations:

- Preprocessing boosts scores
- Preprocessing reduces difference between tokens and types
- TH = 2 tokens
Absolute versus *Normalized Match*

- Intuition: a pair of **short** sentences with $N$ words in common are probably more similar than a pair of **long** sentences with $N$ words in common
- Compensate for sentence length by normalizing w.r.t. total number of types
- Provides scores between 0 and 1
Similarity Measures: Normalized Overlap

\[
MaxOverlap(A, B) = \frac{|A \cap B|}{\max(|A|, |B|)}
\]

\[
MinOverlap(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)}
\]

- \( MaxOverlap(\text{“a great time”}, \text{“a great idea”}) = \frac{2}{3} \)
- \( MinOverlap(\text{“a great time”}, \text{“a great idea”}) = \frac{2}{3} \)

- \( MaxOverlap(\text{“a great time”}, \text{“I had a great idea”}) = \frac{2}{5} \)
- \( MinOverlap(\text{“a great time”}, \text{“I had a great idea”}) = \frac{2}{3} \)
**Results: Normalized Overlap**

**Observations:**

- Normalized overlap not better than absolute overlap
- MaxOverlap and MinOverlap are same with high recall but different with high precision
- While MinOverlap is the conventional overlap coefficient, MaxOverlap works better!
Other Similarity Measures: Jaccard

\[ Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} \]

- Jaccard(“drink beer”, “watch TV”) = 0/4
- Jaccard(“a good time”, “a bad dream”) = 1/5
- Jaccard(“space walk”, “walk space”) = 2/2
- Jaccard(“spam, spam, spam, spam”, “spam”) = 1/1
Similarity Measures: Cosine

\[ \text{cosine}(A,B) = \frac{|A \cap B|}{\sqrt{|A| \times |B|}} \]

- Tends to be more tolerant to differences in sentence length
More similarity measures

- Dice
- Tanimoto
- ...

(see e.g. Manning & Schutze '99)
Results: Other Similarity Measures

Observations:
- Only marginal differences
- Jaccard = Tanimoto
- Tanimoto ≈ cosine

![Graph showing comparison between different similarity measures.](image-url)
Token Weighting

• Intuition: sentences pairs that share **important** words are probably more similar than sentence pairs that share only **unimportant** words

• Using **tf-idf** is the standard approach to term weighting in Information Retrieval (Sparck Jones '72, Salton '73)

• tf-idf a statistical measure to evaluate how important a **term** is to a specific **document** in a document collection

• tf-idf is usually combined with a weighted cosine similarity measure in vector space model
Weighting: Term Frequency

- **term frequency** \((tf)\) indicates how important a term is within a document

\[
    tf(t,d) = \frac{\text{count}(t,d)}{\sum_{m \in d} \text{count}(m,d)}
\]
Weighting: Inverse Document Frequency

- **inverse document frequency** (idf) indicates how important a term is in the total document collection.

\[ \text{idf}(t) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \]

where \(|D|\) is the total no of documents in the collection and \(|\{d \in D : t \in d\}|\) is the no of documents containing term \(t\).
Weighting: tf-idf

- **term frequency - inverse document frequency** \((tf\cdot idf)\)

\[
\text{tf}\_\text{idf}(t,d) = \text{tf}(t,d) \cdot \text{idf}(t)
\]

- *High weight* results from a *local high* frequency (within the document), and *global low* frequency (in the document collection)
- *Low weight* results from *local low* frequency and *overall high* frequency
- Tends to filter out common terms
Weighting: tf-idf for alignment

- Key insight from Nelken & Shieber (EACL06):
  - regard each sentence as a document
  - regard the combined sentences from a pair of texts as a document collection
  - then use cosine similarity with tf-idf weighting for alignment

- Thus
  - If two sentences share a token that does not occur in other sentences, it gets a high weight
  - Conversely, if they share a token that occurs in many other sentences, it gets a low weight
  - Filters out "confusing" words
Results: *tf-idf weighted cosine*

Observations:
- Weighted cosine is somewhat better than plain cosine
Conclusion so far

1. Match on types rather than tokens
2. Preprocess the tokens
3. Normalized measures (MaxSim, Cosine, Jaccard, Dice, Tanimoto) are not better than absolute measures (TypeOverlap)
4. Use tf-idf to weight the tokens
Failed Attempts at Improvement

- Word order may be relevant
  - e.g. "space walk" ≠ "walk space"
  - use token bigram or token trigram?
  - but then "a nice beach" has nothing in common with "a beautiful beach"

- Morphological variations may be irrelevant
  - e.g. "book" and "books"
  - use character bigrams, trigrams, etc.?
  - e.g. \{bo, oo, ok\} and \{bo, oo, ok, ks\}

- 1-to-N alignments are unlikely if N is "large"
  - put upper bound on alignments?

- Weighted variants of other normalized measures
  - E.g. weighted Jaccard
All results

Observations:
• Upper bound on F-score around 60%
• Trade-off between precision and recall is fairly linear
Conclusion

1. Match on types rather than tokens
2. Preprocess the tokens
3. Normalized measures (MaxSim, Cosine, Jaccard, Dice, Tanimoto) are not better than absolute measures (TypeOverlap)
4. Use tf-idf to weight the tokens
5. An F-score of around 60% may be the upper limit for shallow approaches
Discussion

• Many differences are small and may not be statistically significant
• Thresholds may not be stable
  – cross validate
  – ideally, other text material
• Some prediction "errors" may actually be annotation errors
Future work

• Exploit contextual information (Nelken & Shieber '06)
  – Aligned sentences tend to cluster
  – Sequence alignment algorithms (Needleman-Wunsch)
• Try some "expensive" measures
  – Levenshtein distance, Largest Common Subsequence (LCS)
• Machine learning on a combination of measures
• Ultimately, compare to alignment based on deep processing
  – Using knowledge from thesaurus or ontology (e.g. WordNet)
  – Using linguistic knowledge, i.e. as in alignment of parse trees