On Parallel Data Mining

Wang Ling
What is Parallel Data
Parallel Data

Parallel Data/Parallel Corpora are sentences translated in more than 2 languages.
Parallel Data

Parallel Data/Parallel Corpora are sentences translated in more than 2 languages.

| ZH | 天下大勢，分久必合，合久必分。 |

- 三国演义
- Three Kingdoms
Parallel Data

Parallel Data/Parallel Corpora are sentences translated in more than 2 languages.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ZH</strong></td>
<td>天下大勢，分久必合，合久必分。</td>
</tr>
<tr>
<td><strong>EN</strong></td>
<td>Long divided, the world will unite; long united, it will fall apart.</td>
</tr>
<tr>
<td><strong>PT</strong></td>
<td>O império, unir-se-á após a divisão e dividir-se-á após a união</td>
</tr>
</tbody>
</table>
## Parallel Data

Parallel Data/Parallel Corpora are sentences translated in more than 2 languages.

<table>
<thead>
<tr>
<th></th>
<th>ZH</th>
<th>EN</th>
<th>PT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZH</td>
<td>天下大勢，分久必合，合久必分。</td>
<td>Long divided, the world will unite; long united, it will fall apart.</td>
<td>O império, unir-se-á após a divisão e dividir-se-á após a união</td>
</tr>
<tr>
<td>EN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PT</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Parallel Data Characteristics

- Prefer shorter sentence pairs
  - Less expensive inference during alignment
  - Smaller margin of error
Parallel Data Characteristics

● Prefer shorter sentence pairs
  ○ Less expensive inference during alignment
  ○ Smaller margin of error

● More data generally lead to better translations
Parallel Data Characteristics

- Prefer shorter sentence pairs
  - Less expensive inference during alignment
  - Smaller margin of error
- More data generally lead to better translations
- Noisy parallel data is still usable
  - Translation models are robust to noise
Parallel Data in MT

- Parallel Corpora (Training)
- Parallel Corpora (Devel)
- Parallel Corpora (Test)

- Translation Model
- Tuning
- Decoding
- Evaluation
Parallel Data in MT

- Parallel Corpora (Training)
- Parallel Corpora (Devel)
- Parallel Corpora (Test)

- Translation Model
- Tuning
- Decoding
- Evaluation
Parallel Data in MT

Parallel Corpora (Training)

Parallel Corpora (Devel)

Parallel Corpora (Test)

Translation Model

Tuning

Decoding

Evaluation
Parallel Data in MT

- Parallel Corpora (Training)
- Parallel Corpora (Devel)
- Parallel Corpora (Test)

Translation Model

Tuning

Decoding

Evaluation
Parallel Data in MT

Parallel Data is a scarce resource
On Parallel Data Mining
In The Beginning...

...there was no Parallel Corpora
Hansard Corpora (Brown et al., 1990)

- 3 Million Sentences
- French-English
- Parliament Data
Le Québec est encore une fois la cible de fermetures, de coupes et de réductions d'effectifs, 20 ans après l'occupation du bureau de poste de Saint-Clément, qui a mené au moratoire interdisant la fermeture des bureaux de poste ruraux. Depuis, le gouvernement fédéral n'a rien fait pour faire cesser la perte de services essentiels à la population.

En 2011, la moitié des fermetures ou des réductions d'heures était effectuées au Québec. L'an dernier, 17 fermetures sur 31 ont touché le Québec. En avril, cette année, trois comptoirs postaux fermeront leurs portes dans la région de Montréal, sans compter la réduction d'effectifs dans sept autres.
Le Québec est encore une fois la cible de fermetures, de coupes et de réductions d'effectifs, 20 ans après l'occupation du bureau de poste de Saint-Clément, qui a mené au moratoire interdisant la fermeture des bureaux de poste ruraux. Depuis, le gouvernement fédéral n'a rien fait pour faire cesser la perte de services essentiels à la population.

En 2011, la moitié des fermetures ou des réductions d'heures était effectuées au Québec. L'an dernier, 17 fermetures sur 31 ont touché le Québec. En avril, cette année, trois comptoirs postaux fermeront leurs portes dans la région de Montréal, sans compter la réduction d'effectifs dans sept autres.
Le Québec est encore une fois la cible de fermetures, de coupes et de réductions d'effectifs, 20 ans après l'occupation du bureau de poste de Saint-Clément, qui a mené au moratoire interdisant la fermeture des bureaux de poste ruraux. Depuis, le gouvernement fédéral n'a rien fait pour faire cesser la perte de services essentiels à la population.

En 2011, la moitié des fermetures ou des réductions d'heures était effectuées au Québec. L'année dernière, 17 fermetures sur 31 ont touché le Québec. En avril, cette année, trois comptoirs postaux fermeront leurs portes dans la région de Montréal, sans compter la réduction d'effectifs dans sept autres.
Le Québec est encore une fois la cible de fermetures, de coupes et de réductions d'effectifs, 20 ans après l'occupation du bureau de poste de Saint-Clément, qui a mené au moratoire interdisant la fermeture des bureaux de poste ruraux. Depuis, le gouvernement fédéral n'a rien fait pour faire cesser la perte de services essentiels à la population.

En 2011, la moitié des fermetures ou des réductions d'heures était effectuées au Québec. L'an dernier, 17 fermetures sur 31 ont touché le Québec. En avril, cette année, trois comptoirs postaux fermeront leurs portes dans la région de Montréal, sans compter la réduction d'effectifs dans sept autres.

Quebec is once again being targeted for closures, cuts and downsizing. This comes 20 years after the post office in Saint-Clément was occupied, which resulted in a moratorium on the closure of rural post offices. Since then, the federal government has done nothing to stop the loss of essential services.

In 2011, Quebec fell victim to half of the closures and reductions in hours. Last year, 17 of the 31 closures happened in Quebec. In April of this year, three postal outlets in the Montreal area will close, and another seven will make staffing cuts.
Le Québec est encore une fois la cible de fermetures, de coupes et de réductions d'effectifs, 20 ans après l'occupation du bureau de poste de Saint-Clément, qui a mené au moratoire interdisant la fermeture des bureaux de poste ruraux. Depuis, le gouvernement fédéral n'a rien fait pour faire cesser la perte de services essentiels à la population.

En 2011, la moitié des fermetures ou des réductions d'heures était effectuées au Québec. L'an dernier, 17 fermetures sur 31 ont touché le Québec. En avril, cette année, trois comptoirs postaux fermeront leurs portes dans la région de Montréal, sans compter la réduction d'effectifs dans sept autres.

Quebec is once again being targeted for closures, cuts and downsizing. This comes 20 years after the post office in Saint-Clément was occupied, which resulted in a moratorium on the closure of rural post offices. Since then, the federal government has done nothing to stop the loss of essential services.

In 2011, Quebec fell victim to half of the closures and reductions in hours. Last year, 17 of the 31 closures happened in Quebec. In April of this year, three postal outlets in the Montreal area will close, and another seven will make staffing cuts.
Sentence Alignment

Problem: How to obtain sentence level pairs from raw text.

Le Québec est encore une fois la cible de fermetures, de coupes et de réductions d'effectifs, 20 ans après l'occupation du bureau de poste de Saint-Clément, qui a mené au moratoire interdisant la fermeture des bureaux de poste ruraux. Depuis, le gouvernement fédéral n'a rien fait pour faire cesser la perte de services essentiels à la population.

En 2011, la moitié des fermetures ou des réductions d'heures était effectuées au Québec. L'an dernier, 17 fermetures sur 31 ont touché le Québec. En avril, cette année, trois comptoirs postaux fermeront leurs portes dans la région de Montréal, sans compter la réduction d'effectifs dans sept autres.

Quebec is once again being targeted for closures, cuts and downsizing. This comes 20 years after the post office in Saint-Clément was occupied, which resulted in a moratorium on the closure of rural post offices. Since then, the federal government has done nothing to stop the loss of essential services.

In 2011, Quebec fell victim to half of the closures and reductions in hours. Last year, 17 of the 31 closures happened in Quebec. In April of this year, three postal outlets in the Montreal area will close, and another seven will make staffing cuts.
Problem: How to obtain sentence level pairs from raw text.

Le Québec est encore une fois la cible de fermetures, de coupes et de réductions d'effectifs, 20 ans après l'occupation du bureau de poste de Saint-Clément, qui a mené au moratoire interdisant la fermeture des bureaux de poste ruraux.

Quebec is once again being targeted for closures, cuts and downsizing. This comes 20 years after the post office in Saint-Clément was occupied, which resulted in a moratorium on the closure of rural post offices.

Depuis, le gouvernement fédéral n'a rien fait pour faire cesser la perte de services essentiels à la population.

Since then, the federal government has done nothing to stop the loss of essential services.

En 2011, la moitié des fermetures ou des réductions d'heures était effectuées au Québec.

In 2011, Quebec fell victim to half of the closures and reductions in hours.

L'an dernier, 17 fermetures sur 31 ont touché le Québec.

Last year, 17 of the 31 closures happened in Quebec.

En avril, cette année, trois comptoirs postaux fermeront leurs portes dans la région de Montréal, sans compter la réduction d'effectifs dans sept autres.

In April of this year, three postal outlets in the Montreal area will close, and another seven will make staffing cuts.
Sentence Alignment

Problem: How to obtain sentence level pairs from raw text.
Sentence Alignment

Problem: How to obtain sentence level pairs from raw text.

- Sentences are not always translated 1-to-1
Sentence Alignment

Problem: How to obtain sentence level pairs from raw text.

- Sentences are not always translated 1-to-1
- Translations might omit or add sentences
Method (Gale and Church, 1993)

Step 1 : Paragraph Alignment (Manual)

Le Québec est encore une fois la cible de fermetures, de coupes et de réductions d'effectifs, 20 ans après l'occupation du bureau de poste de Saint-Clément, qui a mené au moratoire interdisant la fermeture des bureaux de poste ruraux. Depuis, le gouvernement fédéral n'a rien fait pour faire cesser la perte de services essentiels à la population.

En 2011, la moitié des fermetures ou des réductions d'heures était effectuées au Québec. L'an dernier, 17 fermetures sur 31 ont touché le Québec. En avril, cette année, trois comptoirs postaux fermeront leurs portes dans la région de Montréal, sans compter la réduction d'effectifs dans sept autres.

Quebec is once again being targeted for closures, cuts and downsizing. This comes 20 years after the post office in Saint-Clément was occupied, which resulted in a moratorium on the closure of rural post offices. Since then, the federal government has done nothing to stop the loss of essential services. In 2011, Quebec fell victim to half of the closures and reductions in hours. Last year, 17 of the 31 closures happened in Quebec. In April of this year, three postal outlets in the Montreal area will close, and another seven will make staffing cuts.
Le Québec est encore une fois la cible de fermetures, de coupes et de réductions d'effectifs, 20 ans après l'occupation du bureau de poste de Saint-Clément, qui a mené au moratoire interdisant la fermeture des bureaux de poste ruraux. Depuis, le gouvernement fédéral n'a rien fait pour faire cesser la perte de services essentiels à la population.

En 2011, la moitié des fermetures ou des réductions d'heures était effectuées au Québec. L'an dernier, 17 fermetures sur 31 ont touché le Québec. En avril, cette année, trois comptoirs postaux fermeront leurs portes dans la région de Montréal, sans compter la réduction d'effectifs dans sept autres.

Quebec is once again being targeted for closures, cuts and downsizing. This comes 20 years after the post office in Saint-Clément was occupied, which resulted in a moratorium on the closure of rural post offices. Since then, the federal government has done nothing to stop the loss of essential services.

In 2011, Quebec fell victim to half of the closures and reductions in hours. Last year, 17 of the 31 closures happened in Quebec. In April of this year, three postal outlets in the Montreal area will close, and another seven will make staffing cuts.
Le Québec est encore une fois la cible de fermetures, de coupes et de réductions d'effectifs, 20 ans après l'occupation du bureau de poste de Saint-Clément, qui a mené au moratoire interdisant la fermeture des bureaux de poste ruraux. Depuis, le gouvernement fédéral n'a rien fait pour faire cesser la perte de services essentiels à la population.

Quebec is once again being targeted for closures, cuts and downsizing. This comes 20 years after the post office in Saint-Clément was occupied, which resulted in a moratorium on the closure of rural post offices. Since then, the federal government has done nothing to stop the loss of essential services.
Method (Gale and Church, 1993)

Step 2 : Length-based Sentence Alignment
○ Divide Paragraphs by segment using heuristics

Le Québec est encore une fois la cible de fermetures, de coupes et de réductions d'effectifs, 20 ans après l'occupation du bureau de poste de Saint-Clément, qui a mené au moratoire interdisant la fermeture des bureaux de poste ruraux. Depuis, le gouvernement fédéral n'a rien fait pour faire cesser la perte de services essentiels à la population.

Quebec is once again being targeted for closures, cuts and downsizing. This comes 20 years after the post office in Saint-Clément was occupied, which resulted in a moratorium on the closure of rural post offices. Since then, the federal government has done nothing to stop the loss of essential services.
Depuis, le gouvernement fédéral n'a rien fait pour faire cesser la perte de services essentiels à la population.

Quebec is once again being targeted for closures, cuts and downsizing. This comes 20 years after the post office in Saint-Clément was occupied, which resulted in a moratorium on the closure of rural post offices. Since then, the federal government has done nothing to stop the loss of essential services.
Method (Gale and Church, 1993)

Step 2: Length-based Sentence Alignment

- Divide Paragraphs by segment using heuristics
- Compute segment lengths
Method (Gale and Church, 1993)

Step 2: Length-based Sentence Alignment

- Divide Paragraphs by segment using heuristics
- Compute segment lengths
- Find segment alignments to maximize:

\[ P(\text{match}|\delta), \text{ where } \delta = \frac{l_2 - l_1 c}{\sqrt{l_1 s^2}} \]
Method (Gale and Church, 1993)

\[ \delta = \frac{l_2 - l_1 c}{\sqrt{l_1 s^2}} \]

For EN-FR:
- \( c = 1.06 \)
- \( s = 5.6 \)

For EN-DE:
- \( c = 1.1 \)
- \( s = 7.3 \)
Method (Gale and Church, 1993)

Step 2: Length-based Sentence Alignment

- Divide Paragraphs by segment using heuristics
- Compute segment lengths
- Find segment alignments to maximize:

\[ P(\text{match}|\delta), \text{ where } \delta = \frac{l_2-l_1c}{\sqrt{l_1s^2}} \]
Method (Gale and Church, 1993)

Step 2 : Length-based Sentence Alignment

○ Divide Paragraphs by segment using heuristics
○ Compute segment lengths
○ Find segment alignments to maximize:

\[ P(\text{match}|\delta), \text{ where } \delta = \frac{l_2 - l_1 c}{\sqrt{l_1 s^2}} \]

Le Québec est encore une fois la cible de fermetures, de coupes et de réductions d'effectifs, 20 ans après l'occupation du bureau de poste de Saint-Clément, qui a mené au moratoire interdisant la fermeture des bureaux de poste ruraux.

Quebec is once again being targeted for closures, cuts and downsizing. This comes 20 years after the post office in Saint-Clément was occupied, which resulted in a moratorium on the closure of rural post offices.

Depuis, le gouvernement fédéral n'a rien fait pour faire cesser la perte de services essentiels à la population.

Since then, the federal government has done nothing to stop the loss of essential services.
Method (Gale and Church, 1993)

Results:
- 36/621 (5.8%) alignment errors for EN-FR
- 19/695 (2.7%) alignment errors for EN-DE

<table>
<thead>
<tr>
<th>category</th>
<th>English-French</th>
<th>English-German</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>err</td>
<td>%</td>
</tr>
<tr>
<td>1-0</td>
<td>8</td>
<td>8</td>
<td>100</td>
</tr>
<tr>
<td>1-1</td>
<td>542</td>
<td>14</td>
<td>2.6</td>
</tr>
<tr>
<td>2-1</td>
<td>59</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>2-2</td>
<td>9</td>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td>3-1</td>
<td>1</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>3-2</td>
<td>1</td>
<td>1</td>
<td>100</td>
</tr>
</tbody>
</table>
Related Work
Related Work

- Europarl (Koehn, 2005)
  - European Union Parliament Data
  - 21 European Languages
  - 300K-2000K sentence pairs
Related Work

- United Nations (Rafalovitch and Dale, 2009)
  - Resolution data from the UN General Assembly
  - 6 Languages
  - Average of 3M tokens per language
Web Mining
Web Mining

Previous methods rely on human input to find potential sources of parallel data.
Can we find parallel data automatically from the web?
Web as Parallel Data
(Resnik and Noah, 2002)

Goal: Find parallel web pages and extract the parallel data
Web as Parallel Data
(Resnik and Noah, 2002)

- STRAND - Structural Translation Recognition Acquiring Natural Data (Resnik, 1999)
- Implemented in three steps:
  a. Locate pages that might have translations
  b. Generate pairs of pages that might be parallel
  c. Structural Filtering
STRAND Step 1 - Locate Pages

● Use search engine Altavista's advance search to identify two types of pages:
STRAND Step 1 - Locate Pages

● Use search engine Altavista's advance search to identify two types of pages:
  ○ **Parent pages** contain links to parallel pages
STRAND Step 1 - Locate Pages

- Use search engine Altavista's advance search to identify two types of pages:
  - **Parent pages** contain links to parallel pages
STRAND Step 1 - Locate Pages

- Use search engine Altavista's advance search to identify two types of pages:
  - **Parent pages** contain links to parallel pages
    - Find web pages that contain anchors for at least two languages.
STRAND Step 1 - Locate Pages

- Use search engine Altavista's advance search to identify two types of pages:
  - **Parent pages** contain links to parallel pages
    - Find web pages that contain anchors for at least two languages.
      - (anchor:"english" OR anchor:"anglais" OR anchor:"en") AND
      - (anchor:"french" OR anchor:"francais" OR anchor:"fr")
STRAND Step 1 - Locate Pages

- Use search engine Altavista's advance search to identify two types of pages:
  - **Parent pages** contain links to parallel pages
    - Find web pages that contain anchors for at least two languages.
    - Collect the respective links destinations as parallel candidates
STRAND Step 1 - Locate Pages

- Use search engine Altavista's advance search to identify two types of pages:
  - **Parent pages** contain links to parallel pages
  - **Sibling pages** contain links to a version of the same page in other languages
STRAND Step 1 - Locate Pages

- Use search engine Altavista's advance search to identify two types of pages:
  - **Parent pages** contain links to parallel pages
  - **Sibling pages** contain links to a version of the same page in other languages
STRAND Step 1 - Locate Pages

- Use search engine Altavista's advance search to identify two types of pages:
  - **Parent pages** contain links to parallel pages
  - **Sibling pages** contain links to a version of the same page in other languages
    ■ Find web pages that contain anchors for one language
    (anchor:"english" OR anchor:"anglais" OR anchor:"en")
STRAND Step 1 - Locate Pages

- Use search engine Altavista's advance search to identify two types of pages:
  - **Parent pages** contain links to parallel pages
  - **Sibling pages** contain links to a version of the same page in other languages

- Find web pages that contain anchors for one language
- Collect the sibling and the linking webpages as candidates
STRAND Step 2 - Generate Candidate Pairs

- Generate pairs given the candidates

```
Parent
  Anchor=English
  Anchor=French
  Anchor=German
```

```
Candidate 1
Candidate 2
Candidate 3
```
STRAND Step 2 - Generate Candidate Pairs

- Generate pairs given the candidates

![Diagram showing candidate relationships and anchor languages]

Candidate 1
  
Parent
  
Candidate 2
  
Candidate 3

- English - French
- English - German
- French - German
STRAND Step 2 - Generate Candidate Pairs

- Generate pairs given the candidates

![Diagram showing Sibling (Candidate 1) and Sibling (Candidate 2) with Anchor=French.]

![Diagram showing Candidate 1 and Candidate 2 with English - French.]
STRAND Step 3 - Structural Filtering

- Given a pair of candidates A and B, analyse their HTML structure similarity
STRAND Step 3 - Structural Filtering

- Given a pair of candidates A and B, analyse their HTML structure similarity
STRAND Step 3 - Structural Filtering

- Given a pair of candidates A and B, analyse their HTML structure similarity
STRAND Step 3 - Structural Filtering

- Given a pair of candidates A and B, analyse their HTML structure similarity
STRAND Step 3 - Structural Filtering

- Convert non-markup text into tags with length based feature
STRAND Step 3 - Structural Filtering

- Find best structural alignment (sequence matching)
STRAND Step 3 - Structural Filtering

- Using the alignments, compute the quality of the candidate pair using four features:
STRAND Step 3 - Structural Filtering

- Using the alignments, compute the quality of the candidate pair using four features:
  - $dp$ - difference percentage, ratio between aligned tags and the tag total
STRAND Step 3 - Structural Filtering

- Using the alignments, compute the quality of the candidate pair using four features:
  - $dp$ - difference percentage, ratio between aligned tags and the tag total
  - $n$ - number of aligned text tags of unequal length
STRAND Step 3 - Structural Filtering

- Using the alignments, compute the quality of the candidate pair using four features:
  - $dp$ - difference percentage, ratio between aligned tags and the tag total
  - $n$ - number of aligned text tags of unequal length
  - $c$ - correlation of lengths of aligned text tags
STRAND Step 3 - Structural Filtering

- Using the alignments, compute the quality of the candidate pair using four features:
  - $dp$ - difference percentage, ratio between aligned tags and the tag total
  - $n$ - number of aligned text tags of unequal length
  - $c$ - correlation of lengths of aligned text tags
  - $p$ - significance of $c$
STRAND Step 3 - Structural Filtering

- Using the alignments, compute the quality of the candidate pair using four features:
  - $dp$ - difference percentage, ratio between aligned tags and the tag total
  - $n$ - number of aligned text tags of unequal length
  - $c$ - correlation of lengths of aligned text tags
  - $p$ - significance of $c$

- Classification performed using decision-trees
STRAND Results

● French-English annotated data
  ○ 32 dev samples
  ○ 261 test samples
  ○ 100% Precision
  ○ 69.6% Recall

● Chinese-English annotated data (different implementation)
  ○ 98% Precision
  ○ 61% Recall
Content-based Matching

- STRAND relies only on structure to predict parallel web pages
Content-based Matching

- STRAND relies only on structure to predict parallel web pages
- Can we use the content of the non-markup text to predict parallel web pages?
Content-based Matching

- Given two sentences, calculate their similarity:

  - je vais manger
  - I am going to eat
Content-based Matching

- Given two sentences, calculate their similarity:
  - Compute Viterbi Alignments under an alignment model

```plaintext
je vais manger
I am going to eat
```
Content-based Matching

- Given two sentences, calculate their similarity:
  - Compute Viterbi Alignments under an alignment model
  - Compute Similarity Score:

\[
tsim = \frac{\text{number of alignments}}{\text{number of alignments} + \text{number of unaligned words}} = \frac{3}{5}
\]

je vais manger

I am going to eat
Content VS Structure

- **Setup**
  - 241 annotated samples
  - Ran STRAND step 1 and 2 to generate candidates
  - Compared performance using the 4 STRAND features vs the tsim feature in step 3
Content VS Structure

- **Setup**
  - 241 annotated samples
  - Ran STRAND step 1 and 2 to generate candidates
  - Compared performance using the 4 STRAND features vs the tsim feature in step 3

<table>
<thead>
<tr>
<th></th>
<th>STRAND</th>
<th>tsim</th>
</tr>
</thead>
<tbody>
<tr>
<td>French-English</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>100%</td>
<td>83.3%</td>
</tr>
<tr>
<td>Recall</td>
<td>68.4%</td>
<td>92.1%</td>
</tr>
<tr>
<td>F-Measure</td>
<td>81.2%</td>
<td>87.5%</td>
</tr>
</tbody>
</table>
Related Work

- News Commentary
  - One million words
  - 5 languages
  - News domain
Related Work

- $10^9$ word English-French Parallel Corpora
  - Crawled 40 million files from a variety of websites
  - Generate candidate pairs by parsing the urls
  - Split each web document content into sentences
  - Sentence align document pairs
  - Extracted sentence pairs

<table>
<thead>
<tr>
<th>Year</th>
<th>Dataset</th>
<th>Language</th>
<th>Year</th>
<th>Dataset</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>Hansard</td>
<td></td>
<td>2005</td>
<td>Europarl</td>
<td>EN-FR</td>
</tr>
<tr>
<td>2006</td>
<td></td>
<td></td>
<td>2009</td>
<td>United Nations</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
<td>GIGA EN-FR</td>
<td></td>
</tr>
</tbody>
</table>
Related Work

- Yandex Corpora
  - 1 million sentence pairs
  - Russian-English
Large Scale Web Mining

- The parallel corpora crawl so far has been limited to small domains
  - The Largest Parallel dataset discussed was crawled from 40 million web documents
Large Scale Web Mining

- The parallel corpora crawl so far has been limited to small domains
  - The Largest Parallel dataset discussed was crawled from 40 million web documents
- How can we process the web on a large scale?
Google's Approach
(Uszkoreit et al, 2010)

● Dataset
  ○ 2.5 billion crawled web pages
  ○ 6 languages
Google's Approach (Uszkoreit et al, 2010)

- **Dataset**
  - 2.5 billion crawled web pages
  - 6 languages

- **Problem:** Too many candidate pairs to consider
Google's Approach (Uszkoreit et al, 2010)

- **Dataset**
  - 2.5 billion crawled web pages
  - 6 languages

- **Problem**: Too many candidate pairs to consider

- **Goal**: Develop an *efficient* and *parallelizable* algorithm to filter candidates
Google's Approach (Uszkoreit et al, 2010)

- Cast parallel document candidate generation as a classic IR duplicate detection problem.
  - Translate all documents into English

```
Doc 1 (EN) ───────→ Doc 1 (EN)
Doc 2 (ES) ────→ MT ES-EN ────→ Doc 2 (EN)
Doc 3 (FR) ────→ MT FR-EN ────→ Doc 3 (EN)
Doc 4 (DE) ────→ MT DE-EN ────→ Doc 4 (EN)
        ...          ...          ...          ...
Doc N (FR) ────→ MT FR-EN ────→ Doc N (EN)
```
Google's Approach (Uszkoreit et al, 2010)

- Cast parallel document candidate generation as a classic IR duplicate detection problem.
  - Translate all documents into English
  - Build a inverted-index using the document set

```
Inverted Index
```

```
Extract 5-grams

Doc 1 (EN)
Doc 2 (EN)
Doc 3 (EN)
Doc 4 (EN)
...
Doc N (EN)
```

```
the box
Kungpao Chicken
day after tomorrow
memento mori
```
Google's Approach
(Uszkoreit et al, 2010)

- Cast parallel document candidate generation as a classic IR duplicate detection problem.
  - Translate all documents into English
  - Build a reverse-index using the document set

Inverted Index

- the box → Doc 32 → Doc 92 → Doc 932 → ...
- Kungpao Chicken → Doc 3 → Doc 92 → Doc 653 → ...
- day after tomorrow → Doc 4 → Doc 201 → Doc 781 → ...
- memento mori → Doc 30 → Doc 320 → ...
...
Google's Approach (Uszkoreit et al, 2010)

- Cast parallel document candidate generation as a classic IR duplicate detection problem.
  - Discard entries that contain only documents in one language before translated
Google's Approach (Uszkoreit et al, 2010)

- Cast parallel document candidate generation as a classic IR duplicate detection problem.
  - Discard entries that contain only documents in one language before translated.
Google's Approach (Uszkoreit et al, 2010)

- Cast parallel document candidate generation as a classic IR duplicate detection problem.
  - Discard entries that contain too many documents (>50)
Google's Approach (Uszkoreit et al, 2010)

- Cast parallel document candidate generation as a classic IR duplicate detection problem.
  - Discard entries that contain too many documents (>50)
Google's Approach (Uszkoreit et al, 2010)

- Cast parallel document candidate generation as a classic IR duplicate detection problem.
  - For remaining entries, generate candidate pairs

```
Inverted Index

Kungpao Chicken → Doc 3 → Doc 92 → Doc 653 → ...
```

Google's Approach (Uszkoreit et al, 2010)

- Cast parallel document candidate generation as a classic IR duplicate detection problem.
  - For remaining entries, generate candidate pairs

```
Kungpao Chicken ➔ Doc 3 ➔ Doc 92 ➔ Doc 653 ➔ ...
```

```
Doc 3 ➔ Doc 92
```

```
Doc 3 ➔ Doc 653
```

```
Doc 92 ➔ Doc 653
```

...
Google's Approach (Uszkoreit et al, 2010)

- Cast parallel document candidate generation as a classic IR duplicate detection problem.
  - For remaining entries, generate candidate pairs

```
Doc 3 <-> Doc 92
Doc 3 <-> Doc 653
Doc 92 <-> Doc 653
Doc 34 <-> Doc 92
Doc 2 <-> Doc 426
Doc 202 <-> Doc 991
Doc 201 <-> Doc 312
Doc 111 <-> Doc 983
Doc 92 <-> Doc 653
...```

Google's Approach (Uszkoreit et al, 2010)

- Cast parallel document candidate generation as a classic IR duplicate detection problem.
  - Remove duplicates

![Diagram showing document pairs](image)
Google's Approach (Uszkoreit et al, 2010)

- Cast parallel document candidate generation as a classic IR duplicate detection problem.
  - Remove duplicates

![Diagram showing document pairs](image-url)
Google's Approach
(Uszkoreit et al, 2010)

- Cast parallel document candidate generation as a classic IR duplicate detection problem.
  - Score candidate pairs (bigram-matching)
Google's Approach (Uszkoreit et al, 2010)

- Cast parallel document candidate generation as a classic IR duplicate detection problem.
  - Prune by threshold

![Diagram showing document candidate generation with similarity scores]
Google's Approach
(Uszkoreit et al, 2010)

- Cast parallel document candidate generation as a classic IR duplicate detection problem.
  - Use alignment algorithms to get parallel sentences

```
  Doc 3  Doc 92
  0.6

  Doc 92  Doc 653
  0.4

  Doc 34  Doc 92
  0.9

  Doc 111  Doc 983
  0.6

  Doc 201  Doc 312
  0.5

...```
Google's Approach (Uszkoreit et al, 2010)

- Results
  - Reference of 6818 parallel web pages
Google's Approach (Uszkoreit et al, 2010)

- Results
  - Reference of 6818 parallel web pages
  - Results using only the references:

<table>
<thead>
<tr>
<th>score threshold</th>
<th>0.06</th>
<th>0.10</th>
<th>0.12</th>
<th>0.16</th>
<th>0.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>0.92</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>recall</td>
<td>0.91</td>
<td>0.91</td>
<td>0.90</td>
<td>0.89</td>
<td>0.83</td>
</tr>
</tbody>
</table>
Google's Approach (Uszkoreit et al, 2010)

● Results
  ○ Reference of 6818 parallel web pages

  ○ Results using only the references:

<table>
<thead>
<tr>
<th>score threshold</th>
<th>0.06</th>
<th>0.10</th>
<th>0.12</th>
<th>0.16</th>
<th>0.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>0.92</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>recall</td>
<td>0.91</td>
<td>0.91</td>
<td>0.90</td>
<td>0.89</td>
<td>0.83</td>
</tr>
</tbody>
</table>

  ○ Results using the 2.5 billion web pages+references:

<table>
<thead>
<tr>
<th>score threshold</th>
<th>0.06</th>
<th>0.10</th>
<th>0.12</th>
<th>0.16</th>
<th>0.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>0.88</td>
<td>0.93</td>
<td>0.95</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>recall</td>
<td>0.68</td>
<td>0.65</td>
<td>0.63</td>
<td>0.52</td>
<td>0.38</td>
</tr>
</tbody>
</table>
Google's Approach (Uszkoreit et al, 2010)

- MT results
  - Data

<table>
<thead>
<tr>
<th>Language</th>
<th>baseline</th>
<th>books</th>
<th>web</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>27.5 M</td>
<td>0</td>
<td>271.9 M</td>
</tr>
<tr>
<td>French</td>
<td>479.8 M</td>
<td>228.5 M</td>
<td>4,914.3 M</td>
</tr>
<tr>
<td>German</td>
<td>54.2 M</td>
<td>0</td>
<td>3,787.6 M</td>
</tr>
<tr>
<td>Hungarian</td>
<td>26.9 M</td>
<td>0</td>
<td>198.9 M</td>
</tr>
<tr>
<td>Spanish</td>
<td>441.0 M</td>
<td>15.0 M</td>
<td>4,846.8 M</td>
</tr>
</tbody>
</table>
Google's Approach (Uszkoreit et al, 2010)

- MT results
  - Data
    |          | baseline | books | web      |
    |----------|----------|-------|----------|
    | Czech    | 27.5 M   | 0     | 271.9 M  |
    | French   | 479.8 M  | 228.5 M| 4,914.3 M|
    | German   | 54.2 M   | 0     | 3,787.6 M|
    | Hungarian| 26.9 M   | 0     | 198.9 M  |
    | Spanish  | 441.0 M  | 15.0 M | 4,846.8 M|
  - Scores

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Training Data</th>
<th>WMT 2007 news commentary</th>
<th>WMT 2008 news</th>
<th>WMT 2009 news</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech English</td>
<td>baseline</td>
<td>21.59</td>
<td>14.59</td>
<td>16.46</td>
</tr>
<tr>
<td></td>
<td>web</td>
<td>29.26 (+7.67)</td>
<td>20.16 (+5.57)</td>
<td>23.25 (+6.76)</td>
</tr>
<tr>
<td>German English</td>
<td>baseline</td>
<td>27.99</td>
<td>20.34</td>
<td>20.03</td>
</tr>
<tr>
<td></td>
<td>web</td>
<td>32.35 (+4.36)</td>
<td>23.22 (+2.88)</td>
<td>23.35 (+3.32)</td>
</tr>
<tr>
<td>Hungarian English</td>
<td>baseline</td>
<td>-</td>
<td>10.21</td>
<td>11.02</td>
</tr>
<tr>
<td></td>
<td>web</td>
<td>-</td>
<td>12.92 (+2.71)</td>
<td>14.68 (+3.66)</td>
</tr>
<tr>
<td>French English</td>
<td>baseline</td>
<td>34.26</td>
<td>22.14</td>
<td>26.39</td>
</tr>
<tr>
<td></td>
<td>books</td>
<td>34.73 (+0.47)</td>
<td>22.39 (+0.25)</td>
<td>27.15 (+0.76)</td>
</tr>
<tr>
<td></td>
<td>web</td>
<td>36.65 (+2.39)</td>
<td>23.22 (+1.08)</td>
<td>28.34 (+1.95)</td>
</tr>
<tr>
<td>Spanish English</td>
<td>baseline</td>
<td>43.67</td>
<td>24.15</td>
<td>26.88</td>
</tr>
<tr>
<td></td>
<td>books</td>
<td>44.07 (+0.40)</td>
<td>24.32 (+0.17)</td>
<td>27.16 (+0.28)</td>
</tr>
<tr>
<td></td>
<td>web</td>
<td>46.21 (+2.54)</td>
<td>25.52 (+1.37)</td>
<td>28.50 (+1.62)</td>
</tr>
</tbody>
</table>
End of Parallel Data Web Mining?

- Google can process the whole web with more than 50% recall and 97% precision
End of Parallel Data Web Mining?

- Google can process the whole web with more than \textbf{50\% recall} and 97\% precision
  - Even if we improve recall to 100\%, we will only get twice as much data
End of Parallel Data Web Mining?

- Google can process the whole web with more than **50% recall** and 97% precision:
  - Even if we improve recall to 100%, we will only get twice as much data
  - Adding 1000M sentence pairs to 1M yields a vast improvement
End of Parallel Data Web Mining?

- Google can process the whole web with more than 50% recall and 97% precision:
End of Parallel Data Web Mining?

- Google can process the whole web with more than **50% recall** and 97% precision:
  - Even if we improve recall to 100%, we will only get twice as much data
  - Adding 1000M sentence pairs to 1M yields a vast improvement
  - But adding 1000M to another 1000M will not...
End of Parallel Data Web Mining?

Google can process the whole web with more than **50% recall** and 97% precision:

- Even if we improve recall to 100%, we will only get twice as much data.
- Adding 1000M sentence pairs to 1M yields a vast improvement.
- But adding 1000M to another 1000M will not...
- **Is this the end of the line?**
End of Parallel Data Web Mining?

- The Web is evolving...

The Intelligence is in the Connections
End of Parallel Data Web Mining?

• The Web is evolving...
  ○ The Web 1.0 era has past
End of Parallel Data Web Mining?

- The Web is evolving...
  - The Web 1.0 era has past
  - The emergence of new media is radically changing the form information is spreading
    - Blogs (WordPress)
    - Wikis (Wikipedia)
    - Social Networks (Facebook)
    - Microblogs (Twitter)
End of Parallel Data Web Mining?

- The Web is evolving...
  - The Web 1.0 era has past
  - The emergence of new media is radically changing the form information is spreading
    - Blogs (WordPress)
    - Wikis (Wikipedia)
    - Social Networks (Facebook)
    - Microblogs (Twitter)
  - Web Mining Methods must also adapt and innovate to accommodate the emerging new Web
Microblogs as Parallel Corpora
Microblog Translation

- Posts in Microblogs (Twitter) and Social Networks (Facebook) tends to be **noisier** and more **informal** than previously seen media
Microblog Translation

- Posts in Microblogs (Twitter) and Social Networks (Facebook) tends to be noisier and more informal than previously seen media

msg 4 Warren G his cday is today 1 yr older
Microblog Translation

- Posts in Microblogs (Twitter) and Social Networks (Facebook) tends to be noisier and more informal than previously seen media
  - Jargon

msg 4 Warren G his cday is today 1 yr older
Microblog Translation

- Posts in Microblogs (Twitter) and Social Networks (Facebook) tends to be **noisier** and more **informal** than previously seen media
  - Jargon
  - Abbreviations

**msg 4** Warren G his cday is today 1 yr older
Microblog Translation

- Posts in Microblogs (Twitter) and Social Networks (Facebook) tends to be **noisier** and more **informal** than previously seen media
  - Jargon
  - Abbreviations
  - Orthographic errors

happy singles day in China - sorry I won't be celebratin witchu
Microblog Translation

- Posts in Microblogs (Twitter) and Social Networks (Facebook) tend to be **noisier** and more **informal** than previously seen media
  - Jargon
  - Abbreviations
  - Orthographic errors
  - Syntactic errors

I loving being adventurous! ;D
Microblog Translation

- Posts in Microblogs (Twitter) and Social Networks (Facebook) tends to be **noisier** and more **informal** than previously seen media
  - ...  
  - Orthographic errors  
  - Syntactic errors  
  - Emoticons

I loving being adventurous! ;D
### Microblog Translation

- **Problem:** Current parallel corpora are generally **clean and formal.**

<table>
<thead>
<tr>
<th>EUROPARL</th>
<th>Microblogs (Sina Weibo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please rise, then, for this minute’s silence.</td>
<td>msg 4 Warren G his cday is today 1 yr older. happy cday may god bless u and the....</td>
</tr>
<tr>
<td>Madam president, on a point of order.</td>
<td>happy singles day in China - sorry I won't be celebratin witchu, I have my lovel...</td>
</tr>
<tr>
<td>Thank you, mr Segni, I shall do so gladly.</td>
<td>Skydiving was incredible! Such an amazing feeling! I loving being adventurous! ;D</td>
</tr>
</tbody>
</table>
Microblog Translation

- Problem: Current parallel corpora are generally **clean and formal**. But Microblogs are **noisy and informal**.

<table>
<thead>
<tr>
<th>EUROPARL</th>
<th>Microblogs (Sina Weibo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please rise, then, for this minute’s silence.</td>
<td>msg 4 Warren G his cday is today 1 yr older. happy cday may god bless u and the....</td>
</tr>
<tr>
<td>Madam president, on a point of order.</td>
<td>happy singles day in China - sorry I won't be celebratin witchu, I have my lovel...</td>
</tr>
<tr>
<td>Thank you, mr Segni, I shall do so gladly.</td>
<td>Skydiving was incredible! Such an amazing feeling! I loving being adventurous! ;D</td>
</tr>
</tbody>
</table>
What’s the problem with current MT

msg 4 Warren G his cday is today 1 yr older.
What’s the problem with current MT

msg 4 Warren G his cday is today 1 yr older.

Bing Translator
What’s the problem with current MT

msg 4 Warren G his cday is today 1 yr older.
What’s the problem with current MT

msg 4 Warren G his cday is today 1 yr older.

Bing Translator

味精 4 沃伦 G 他 cday 今天是较旧的 1 年。

monosodium glutamate 4 warren G he cday today is older 1 year.
Microblog Translation

- Previous methods attempt to amortize this problem by sanitizing/cleansing the input
  - Mostly by replacing abbreviations with expanded forms ("u" to "you") or correct orthographic errors
  - But, this does not solve all problems
Microblog Translation

- Previous methods attempt to amortize this problem by sanitizing/cleansing the input
  - Mostly by replacing abbreviations with expanded forms("u" to "you") or correct orthographic errors
  - But, this does not solve all problems
- How can we get parallel data in this domain?
Microblog Mining

- ...and we found this

Justin Liang–Te Chiu
November 2 near Pittsburgh

Achievement: Failed Driver License Test 3 times in a row. What's better, I am making progress every time. Let's see how far we can go.

考駕照連續失敗第三次，不過每次都有進步。這次我已經不會生氣了，以修身養性來說也是不錯的過程。

Like · Comment

4 people like this.

Yubin Kim Aww, I'm sorry to hear that! I took my test very recently and there's some tricks that they catch you on that I can tell you about.
November 2 at 9:19pm · Like

Write a comment...
Microblog Mining

- Does this also happen in Microblogs?
Microblog Mining

Does this also happen in Microblogs?

Skydiving was incredible! Such an amazing feeling! I loving being adventurous! ;D

Meeting Yao Ming for the first time! So great to be back in China for the Mission Hills World Celebrity Pro-Am. Will post pictures soon! 第一次和姚明见面!又回到中国的感觉太棒了!这次是为观澜湖世界名人赛。照片稍等片后! Thanks.
Microblog Mining

● Does this also happen in Microblogs?

"I am the light and I am the dark. And beyond the light and the dark, I am and God is." 我是 光明, 我也是黑暗。超越光明和黑暗, 我 是, 神是。

msg 4 Warren G his cday is today 1 yr older. happy cday may god bless u and the... - 发信息给  Warren G，今天是他的生日, 又 老了一岁了。生日快乐, 愿上帝保佑你和 ...
Microblog Mining

Does this also happen in Microblogs?

Summer Stand, The Drenched Show 2012
2012 싸이 훨씬 더 희뽑 쇼

進進進進进球了! 克罗斯为拜仁破门！
Toooooooooooooooooool 6:1! Kroos trifft für den FC Bayern!
Microblog Translation

• But there is a catch...
Microblog Translation

- But there is a catch...
  - Not all multilingual tweets are parallel

[GD's Twitter] ONE OF A KIND 的 M/V 马上 就要公开了 !! Y’all Ready for this ?呃啊 啊啊, 好紧张啊~还请大家多多支持！

转发微博《南方小羊牧场》 11月9号北 美上映。Showtime is coming up soon...
Microblog Translation

● But there is a catch...
  ○ Not all multilingual tweets are parallel
  ○ Finding the parallel segments in the message is not trivial
● But there is a catch...
  ○ Not all multilingual tweets are parallel
  ○ Finding the parallel segments in the message is not trivial

I wanna be here every year if possible~!

可能的话，我想每年来这里～！
就算像有 的人担心的那样我只是“昙花一现"，我还是会非常努力 的 ...
因为回头看的话，一切 都只是一 瞬的 ...

...^^
Microblog Translation

- But there is a catch...
  - Not all multilingual tweets are parallel
  - Finding the parallel segments in the message is not trivial

I wanna be here every year if possible~! I may worry about me being "ephemeral" but I'll try my best, ...^^

可能的话，我想每年来这里～！就算有人担心我只是“昙花一现”，我还会非常努力的...^^

因为回头看的话，一切都只是一瞬的...^^
Microblog Translation

- Comparison with previous work
Microblog Translation

- Comparison with previous work
  - No structure in posts
    - Structural Filtering is not possible
Microblog Translation

- Comparison with previous work
  - No structure in posts
    - Structural Filtering is not possible
  - Content-based matching is ok
    - But, previous work assumes that a pair of documents will be given
Microblog Alignment Model

- Solution: Consider all spans for matching
Microblog Alignment Model

- Solution: Consider all spans for matching

1) Select a left span \((p,q)\) and right span \((u,v)\)

left span: 一起努力吧。 // @tag: We fighting together
right span: p=0 \hspace{2cm} q=4 \hspace{2cm} u=10 \hspace{2cm} v=12

2) Select languages for spans \((l,r)\)

l=cn \hspace{2cm} r=en

3) Generate alignments \((a)\) from left to right spans

\[ a=\{(1,11),(2,12)\} \]
Microblog Alignment Model

- **Solution:** Consider all spans for matching
- **Problem:** Running the Viterbi Alignments for all possible spans is intractable
Microblog Alignment Model

- **Solution:** Consider all spans for matching
- **Problem:** Running the Viterbi Alignments for all possible spans is intractable
  - **Answer:** Dynamic Programming
  - Reuse Viterbi Alignments for previously processed spans
Viterbi Alignments IBM1

Hoje  Eu  fui  a  Espanha

I  went  to  Spain
I went to Spain

Hoje → Eu → fui → a → Espanha

I → went → to → Spain

0.1 0.7 0.1 0.2 0
I went to Spain

Score = 0.7
Viterbi Alignments IBM1

Score = 0.7 * 0.5
Viterbi Alignments IBM1

Score = 0.7 * 0.5 * 0.4 * 0.9

Hoje → Eu → fui → a → Espanha

I → went → to → Spain
Viterbi Alignments IBM1

Score = 0.7 * 0.5 * 0.4 * 0.9

Number of operations : 5 * 4
Viterbi Alignments IBM1 (add target)

Score = 0.7 * 0.5 * 0.4 * 0.9

Today I went to Spain
Viterbi Alignments IBM1 (add target)

Score = 0.7 * 0.5 * 0.4 * 0.9
Viterbi Alignments IBM1 (add target)

Score = 0.3 * 0.7 * 0.5 * 0.4 * 0.9
Viterbi Alignments IBM1 (add target)

Score = $0.3 \times 0.7 \times 0.5 \times 0.4 \times 0.9$

Number of operations : 5
Viterbi Alignments IBM1 (rm target)

Score = 0.3 * 0.7 * 0.5 * 0.4 * 0.9

Today - I - went - to - Espanha

Hoje - Eu - fui - a - Espanha

0.3 0.7 0.5 0.4 0.9
Viterbi Alignments IBM1 (rm target)

Score = \(0.3 \times 0.7 \times 0.5 \times 0.4\)

Number of operations : 1
Viterbi Alignments IBM1 (add source)

Score = 0.3 * 0.7 * 0.5 * 0.4

Today I went to Espanha :)
Viterbi Alignments IBM1 (add source)

Score = 0.3 * 0.7 * 0.5 * 0.4

Hoje → Eu → fui → a → Espanha → :)
Viterbi Alignments IBM1 (add source)

Score = 0.3 * 0.7 * 0.5 * 0.4

Number of operations : 4
Viterbi Alignments IBM1 (rm source)

Score = 0.3 * 0.7 * 0.5 * 0.4

Today I went to Espanha :)

Eu fui a Espanha
Viterbi Alignments IBM1 (rm source)

Score = 0.3 * 0.7 * 0.5 * 0.4
Viterbi Alignments IBM1 (rm source)

Score = 0.1 * 0.7 * 0.5 * 0.4

Number of operations : 5
Viterbi Alignments IBM1 (rm source)

Score = 0.3 * 0.7 * 0.5 * 0.4

Worst case: recompute alignments

Today  I  went  to

Hoje  Eu  fui  a  Espanha  :)

\[\text{Score} = 0.3 \times 0.7 \times 0.5 \times 0.4\]

Worst case: recompute alignments
Viterbi Alignments IBM1 (rm source)

Score = 0.3 * 0.7 * 0.5 * 0.4

Worst case: recompute alignments

Today I went to Espanha :)

0.3  0.7  0.5  0.4
Viterbi Alignments IBM1

- Add target = $O(S)$
- Rm target = $O(1)$
- Add source = $O(T)$
- Rm source = $O(S\times T)$
Viterbi Alignments IBM1

- Add target = $O(S)$
- Rm target = $O(1)$
- Add source = $O(T)$
- Rm source = $O(S \times T)$
Microblog Alignment Model

- **Solution**: Consider all spans for matching
- **Problem**: Running the Viterbi Alignments for all possible spans is intractable
  - **Answer**: Dynamic Programming
  - Reuse Viterbi Alignments for previously processed spans
  - Reduces Complexity from $O(N^6)$ to $O(N^4)$
Results

● Dataset
  ○ Crawled 65 million targeted tweets from Sina Weibo
  ○ Used 1.6 billion tweets GardenHose (Twitter)
  ○ Reporting results for Mandarin-English
MT Results

- Training Parallel Data
  - Weibo
    - Approximately 1M multilingual tweets from Sina Weibo
    - Expect 337K parallel sentences
    - Microblog Domain
MT Results

- Training Parallel Data
  - **FBIS dataset**
    - 300K parallel sentences
    - News Domain
MT Results

- Training Parallel Data
  - **FBIS dataset**
    - 300K parallel sentences
    - News Domain
  - **NIST dataset**
    - 8M parallel sentences (including FBIS)
    - News Domain
MT Results

- Development and Test sets
  - Weibo
    - Built by annotating weibo tweets manually
    - 1000 dev
    - 1000 test
    - Microblog domain
MT Results

- Development and Test sets
  - **Syndicate**
    - Extracted from project syndicate (Parallel website)
    - 1000 dev
    - 1000 test
    - News and political domain
    - Up-to-date
MT Results

- Baseline

<table>
<thead>
<tr>
<th></th>
<th>Weibo</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZH-EN</td>
<td>EN-ZH</td>
</tr>
<tr>
<td>FBIS</td>
<td>10.4</td>
</tr>
</tbody>
</table>
MT Results

- Baseline
  - More Data = Better

<table>
<thead>
<tr>
<th></th>
<th>Weibo</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZH-EN</td>
<td>10.4</td>
</tr>
<tr>
<td>EN-ZH</td>
<td>12.3</td>
</tr>
<tr>
<td>FBIS</td>
<td>11.4</td>
</tr>
<tr>
<td>NIST</td>
<td>13.9</td>
</tr>
</tbody>
</table>
## MT Results

- **In-domain data**

  Significant improvements (30-40%) on microblogs (in-domain)

<table>
<thead>
<tr>
<th></th>
<th>ZH-EN</th>
<th>EN-ZH</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FBIS</strong></td>
<td>10.4</td>
<td>12.3</td>
</tr>
<tr>
<td><strong>NIST</strong></td>
<td>11.4</td>
<td>13.9</td>
</tr>
<tr>
<td><strong>Weibo</strong></td>
<td><strong>15.7</strong></td>
<td><strong>17.2</strong></td>
</tr>
</tbody>
</table>
**MT Results**

- In-domain data

  Even More Data = Even Better

<table>
<thead>
<tr>
<th></th>
<th>Weibo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZH-EN</td>
</tr>
<tr>
<td>FBIS</td>
<td>10.4</td>
</tr>
<tr>
<td>NIST</td>
<td>11.4</td>
</tr>
<tr>
<td>Weibo</td>
<td>15.7</td>
</tr>
<tr>
<td>FBIS+Weibo</td>
<td>16.5</td>
</tr>
<tr>
<td>NIST+Weibo</td>
<td><strong>16.9</strong></td>
</tr>
</tbody>
</table>
MT Results

- Out of domain
  - News data is better but still benefits from crawled data

<table>
<thead>
<tr>
<th></th>
<th>Syndicate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZH-EN</td>
</tr>
<tr>
<td>FBIS</td>
<td>9.4</td>
</tr>
<tr>
<td>NIST</td>
<td>11.5</td>
</tr>
<tr>
<td>Weibo</td>
<td>8.8</td>
</tr>
<tr>
<td>FBIS+Weibo</td>
<td>11.7</td>
</tr>
<tr>
<td>NIST+Weibo</td>
<td><strong>13.3</strong></td>
</tr>
</tbody>
</table>
New Translations (in English)?
New Translations (in English)?

谢谢=thx, 你=u

To Colton Lopez, thx for the love! 对 Colton Lopez说，谢谢你的爱

have u ever really lived in beijing？ 你是否真的住过北京
New Translations (in English)?

To Colton Lopez, thx for the love!

Google Translate

要科尔顿·洛佩兹，thx的爱！

have u ever really lived in beijing?

Google Translate

具有u曾经真的住在北京吗？
New Translations (in Chinese)?
New Translations (in Chinese)?

TMD=damn, TM=damn
New Translations (in Chinese)?

TMD=damn, TM=damn

他妈的—Ta Ma De
New Translations (in Chinese)?

TMD=damn, TM=damn

Life is like the game "Angry Birds". When you fail, there are always some damn stupid pigs laughing at you.

人生就像"愤怒的小鸟", 当你失败时, 总有TMD几只笨猪在笑
New Translations (in Chinese)?

Life is like "Angry Birds", when you fail, there are always a few bungee laughing TMD
New Translations (in Chinese)?

囧=embarrassed
New Translations (in Chinese)?

囧 = embarrassed

I'm so embarrassed. 我囧死了。
New Translations (in Chinese)?

我囧死了。

Google Translate

I embarrassing dead.
New Translations (in Chinese)?

屌丝 = loser
Today I heard a male foreign loser roaring in anger on the phone, "You are a liar! You don't love me at all! All you want to do is practise oral English!!!

今天在地铁站,看到一个外国男屌丝在电话咆哮:你是 个骗子!你一点都不爱我!你只是想和我练口语!
New Translations (in Chinese)?

Today, in the subway station, saw a foreign man 屌丝在电话咆哮

Google Translate

Today, in the subway station, saw a foreign man on the phone roar Cock wire
Microblog Parallel Data

- utopia
  - 3M Mandarin-English sentence pairs from Weibo
  - 150-200K sentence pairs for 6 pairs from Twitter
Thx y'all for ur attention ;)}