COLESIR at CLEF 2007: from English to French via Character N-Grams

Jesús Vilares
Computer Science Dept.
University of A Coruña
jvilares@udc.es

Michael P. Oakes
School of Computing and Technology
University of Sunderland
Michael.Oakes@sunderland.ac.uk

Manuel Vilares
Computer Science Dept.
University of Vigo
vilares@uvigo.es
Index

- Introduction
- Previous approaches
- Our proposal
- Evaluation
- Conclusions and future work
Index

- Introduction
- Previous approaches
- Our proposal
- Evaluation
- Conclusions and future work
Translation in CLIR

- Techniques of **Machine Translation (MT)**
  - Softened restrictions
    - Not limited to just one translation
    - Not limited by syntax

- **Conventional MT tools** (e.g., SYSTRAN)
  - Single well-formed translation
  - Dismisses advantages of MT in CLIR
Translation in CLIR (cont.)

- **Bilingual dictionaries**
  - Problems with out-of-vocabulary words (misspellings, unknown words)
  - Normalization
  - Word-Sense Disambiguation (WSD)

- **Parallel corpora**
  - Automatic generation of dictionaries:
    - Collocations
    - Association measures
  - Probabilistic translation measure
  - No normalization
Character $N$-Grams

tomatoes $^{n=5}$ $\rightarrow \{ \text{-tomat-}, \text{-omato-}, \text{-matoe-}, \text{-atoes-} \}$

Applications:
- Language recognition
- Misspelling processing
- Information Retrieval
  - Reduction of vocabulary size (dictionary)
  - Asian languages (no delimiters)
Index

- Introduction
- Previous approaches
- Our proposal
- Evaluation
- Conclusions and future work
No word normalization

Language-independent:
- No language-specific processing
- Applicable to very different languages

Knowledge-light approach:
- Minimal linguistic information and resources

Robustness:
- Out-of-vocabulary words
INDEXING

SPLITTING MODULE

INDEXING ENGINE

INDEX

(en)

tomatoes

tomat- -omato-
-matoe- -atoes

J. Vilares, M. P. Oakes and M. Vilares. From English to French via Character N-Grams-- p. 8
McNamee and Mayfield, 2004 (cont.)

STANDARD WORD-LEVEL TRANSLATION

INDEXING

N-GRAM DIRECT TRANSLATION

QUERY (es)

tomates

tomatoes

tomat--omato--matoe--atoes

INDEX

QUERY (es)

tomates

tomat--omato--matoe--atoes

INDEXING ENGINE

RETRIEVAL ENGINE

SPLITTING MODULE

SPLITTING MODULE

SPLITTING MODULE

TRANSLATION ENGINE

TRANSLATION ENGINE

RETRIEVAL ENGINE

J. Vilares, M P. Oakes and M. Vilares. From English to French via Character N-Grams-- p. 8
**N-Gram Alignment Algorithm**

- **Input:** parallel corpus aligned at paragraph-level
  - Text splitted into \( n \)-grams

- **Process:** for each source \( n \)-gram of source language:
  1. To locate source language paragraphs containing it
  2. To identify parallel paragraphs in target language
  3. To calculate **translation score** for each \( n \)-gram in target paragraphs (**ad-hoc association measure**).
  4. **Potential translation:** target \( n \)-gram with highest score.

- **Output:** \( n \)-gram-level alignment
Drawbacks:

- **Very slow** (several days): *not accurate for testing*
- Single translation
Index

- Introduction
- Previous approaches
- **Our proposal**
- Evaluation
- Conclusions and future work
Goals

- Testing tool
- To speed up the training process
- Multiple translations
- Freely available resources
  - More transparency
  - Reduce effort
Differences

- **Freely available resources:**
  - Parallel corpus: **EUROPARL** *(Koehn, 2005)*
  - Statistical aligner: **GIZA++** *(Och and Ney, 2003)*
  - Retrieval engine: **TERRIER** *(http://ir.dcs.gla.ac.uk/terrier/)*

- **Standard association measures:**
  - Dice coefficient
  - Mutual Information
  - Log-likelihood

- **Alignment in two phases:**
  1. Word-level alignment
  2. $N$-gram-level alignment
**N-Gram Alignment Algorithm**

**Input:** parallel corpus aligned at paragraph-level

**Process:** two phases

1. **Word-level alignment** using GIZA++ (slowest): filtering

2. **N-gram-level alignment:**
   - Aligned words as *weighted word-level parallel corpus*
   - **Association measures** between cooccurring *n*-grams
     - Likelihood of cooccurrences weighted according to their alignment probabilities (from word-level alignment)

**Output:** *n*-gram-level alignment
Optimizations:

- **Input word-translation probability threshold** $W$ ($W=0.15$)
  - Input word pairs / output $n$-gram pairs: $\sim 95\%$ reduction

- **Bidirectional word alignment** ($EN2FR \cap FR2EN$)
  - Input word pairs / output $n$-gram pairs: $\sim 50\%$ reduction
Index

- Introduction
- Previous approaches
- Our proposal
- Evaluation
- Conclusions and future work
Evaluation

- **English-to-French** run \((EN2FR)\)
- 4-grams \((McNamee and Mayfield, 2004)\)
- **TERRIER** retrieval engine: DFR paradigm
- InL2 weight

**Corpus:** CLEF 2007 robust track \((Cross-Language Evaluation Forum)\)

<table>
<thead>
<tr>
<th>collection (FR)</th>
<th>size</th>
<th>#docs.</th>
<th>#topics (EN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeMonde 94 + SDA 94</td>
<td>243 MB</td>
<td>87,191</td>
<td>100 ((training))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100 ((test))</td>
</tr>
</tbody>
</table>
Querying

- title + description topic fields

Querying process:

- Split source language query into $n$-grams
- Replaced by their $N$ highest scored aligned target n-grams:
  - Tuned using English-to-Spanish experiments (EN2ES)
    - Dice coefficient $N=1$
    - Mutual Information $N=10$
    - Log-likelihood $N=1$
- Submit translated query
Precision vs. Recall

TRAINING set

TEST set

J. Vilares, M. P. Oakes and M. Vilares. From English to French via Character N-Grams—p. 19
Precision at top $D$ documents

![Graph showing Precision vs Documents retrieved for different languages and metrics.](image)

**TRAINING set**

**TEST set**

J. Vilares, M. P. Oakes and M. Vilares. From English to French via Character N-Grams– p. 20
Index

- Introduction
- Previous approaches
- Our proposal
- Evaluation
- Conclusions and future work
Conclusions

- CLIR using \textit{n-grams as indexing and translation units}

- \textit{N-gram alignment in two phases: speeds up process}
  1. Word-level alignment (\textit{concentrates complexity})
  2. \textit{N}-gram-level alignment

- \textbf{Optimizations} during \textit{word-level alignment}:
  - Word-translation probability threshold
  - Bidirectional alignment

- \textbf{Dice and log-likelihood} perform better
Future work

- New languages
- Remove diacritics
- Remove stopwords and/or *stopngrams* (obtained automatically)
- Simplify word-level alignment (*bottleneck*)
- **Direct evaluation of *n*-gram alignments**
The End

www.grupocole.org

Go back to the beginning of the presentation
### N-Gram Contingency Table

For a pair of n-grams $(u, v)$, the contingency table can be represented as:

<table>
<thead>
<tr>
<th></th>
<th>$V=v$</th>
<th>$V\neq v$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U=u$</td>
<td>$O_{11}$</td>
<td>$O_{12}$</td>
</tr>
<tr>
<td>$U\neq u$</td>
<td>$O_{21}$</td>
<td>$O_{22}$</td>
</tr>
</tbody>
</table>

- $O_{11}$: Number of occurrences where both $U$ and $V$ are $u$ and $v$ respectively.
- $O_{12}$: Number of occurrences where $V$ is $v$ and $U$ is not $u$.
- $O_{21}$: Number of occurrences where $U$ is not $u$ and $V$ is $v$.
- $O_{22}$: Number of occurrences where both $U$ and $V$ are not $u$ and not $v$ respectively.

The table is completed with:

- $R_1 = O_{11} + O_{12}$
- $R_2 = O_{21} + O_{22}$
- $C_1 = O_{11}$
- $C_2 = O_{21}$
- $N = R_1 + R_2$
The likelihood of a cooccurrence is inherited from the probability of its containing word alignment:

\[ P(ngram_{iu} \rightarrow ngram_{jv}) = P(word_u \rightarrow word_v) \]

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>tomate</td>
<td>tomato</td>
<td>0.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>↓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tomat-</td>
<td>-omate</td>
<td>tomat-</td>
<td>-omato</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>↓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tomat-</td>
<td></td>
<td>tomat-</td>
<td></td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tomat-</td>
<td>-omato</td>
<td></td>
<td></td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-omate</td>
<td>tomat-</td>
<td></td>
<td></td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-omate</td>
<td>-omato</td>
<td></td>
<td></td>
<td>0.80</td>
</tr>
</tbody>
</table>
Also reflected in the contingency table. E.g.:

\[ O_{11}(ngram_{iu}, ngram_{jv}) = \sum_{uk/ngram_{iu} \in N(word_{uk})} P(word_{uk} \rightarrow word_{vk}) \]

<table>
<thead>
<tr>
<th>tomate</th>
<th>tomato</th>
<th>0.80</th>
</tr>
</thead>
<tbody>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>tomatitos</td>
<td>tomatoes</td>
<td>0.65</td>
</tr>
<tr>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>tomat-</td>
<td>tomat-</td>
<td>1.45 = 0.80 + 0.65</td>
</tr>
</tbody>
</table>
N-Gram Association Measures

- **Dice Coefficient:**

\[
\text{Dice}(g_s, g_t) = \log \frac{2O_{11}}{R_1 + C_1}
\]

- **Mutual Information:**

\[
\text{MI}(g_s, g_t) = \log \frac{NO_{11}}{R_1 C_1}
\]

- **Log-likelihood:**

\[
\log l(g_s, g_t) = 2 \sum_{i,j} O_{ij} \log \frac{NO_{ij}}{R_i C_j}
\]