Using Surface Syntactic Parser & Deviation from Randomness

Jean-Pierre Chevallet IPAL I2R
Gilles Sérasset CLIPS IMAG
Outline

• Monolingual track
  – French, Russian, Finnish
  – Deviation from randomness

• Bilingual track
  – Bilingual Association Thesaurus for disambiguating Query Translation
Goal of Monolingual experiment

- Compare Deviation from randomness Weighting model, against some other
  - nnn, bnn, lnc, ntc, ltc, atn, dtn, Okapi
- Learn the best parameters
- Use surface syntactic parsing
  - For all documents and queries
  - Ensure correct linguistic stemming
  - Correct split of glued words
- A test for the XIOTA, XML IR system
Global Schema of treatment

- Raw Corpus
  - Parser
  - Corpus Tagged
  - Stem + SW
    - Vector Weighting

- Queries
  - Parser
  - Queries Tagged
    - Association Thesaurus
      - Bilingual Dic.
    - Matching
Deviation from Randomness

- Probabilistic model
- Compute the deviation of statistical repartition of term from a random distribution
- Formula take into account, corpus size and document size
- Only one constant \( c \) : weight normalization for the document length compared to the average length
Influence of C Value in DFR

Precision with C (stem AD)

Precision %

C values (linear)
Influence of $c$ in DFR

Precision with $C$ (stem AD)
Influence of query weighting in Finnish collection
Comment on results

- Deviation from Randomness is very stable under query weighing, and is the best weighting on French and Finnish: we use it for 2004
- Good values of c between 0.6 and 1.0
- When using syntactic parsing
  - No need to a stop list: using grammatical categories
  - Stemming is done, with word splitting
- But these curves use classical stemming and stop list ...(data from Savoy for Finnish)
Results for 2004 Mono Lingual

Russian: 35%
Results for 2004 Mono Lingual

French : 44 %
Results for 2004 Mono Lingual

Finnish : 53%
Comments on results

- Use of parsing for all 3 languages
- Best absolute results on Finnish
  - Results are better than our training in 2003
  - This is an agglutinative language, in our training we have not used the syntactic parsing
- Results in French are lower than our training
  - We have used both parsing and stemming + stop list to recover possible parsing errors
- There is still a lot on query under 10% of precision
  - We should examine closely why we cannot solve these queries: we probably need additional data like good thesaurus or dedicated knowledge base
Topic Translation

- Translation of query vectors
- Building bilingual dictionaries available at CLIPS and online
- French and English as topic language
- Russian and Finnish use Logos web site but only for terms in the topic
- All bilingual dictionary in the same XML file type
Multilingual Experiments

- Construction of the dictionaries

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>nb entries</th>
<th>avg nb trans</th>
<th>max nb trans</th>
</tr>
</thead>
<tbody>
<tr>
<td>fr - en</td>
<td>21417</td>
<td>1.92417</td>
<td>22</td>
</tr>
<tr>
<td>fr - fi</td>
<td>791</td>
<td>1.06574</td>
<td>4</td>
</tr>
<tr>
<td>fr - ru</td>
<td>604</td>
<td>1.06126</td>
<td>3</td>
</tr>
<tr>
<td>en - fr</td>
<td>24542</td>
<td>1.67916</td>
<td>25</td>
</tr>
<tr>
<td>en - fi</td>
<td>867</td>
<td>1.11649</td>
<td>5</td>
</tr>
<tr>
<td>en - ru</td>
<td>15331</td>
<td>2.09901</td>
<td>30</td>
</tr>
</tbody>
</table>
Size of Bilingual Dictionaries

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Nb of Entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>fr - ru</td>
<td>604</td>
</tr>
<tr>
<td>fr - fi</td>
<td>791</td>
</tr>
<tr>
<td>en - fi</td>
<td>867</td>
</tr>
<tr>
<td>en - ru</td>
<td>15331</td>
</tr>
<tr>
<td>fr - en</td>
<td>21417</td>
</tr>
<tr>
<td>en - fr</td>
<td>24542</td>
</tr>
</tbody>
</table>
Topic Translation

• Substitute each terms by all available translation
• Divide the weight of each translation by the number of translation
• Selection of some better translation by filtering using an association thesaurus
Multilingual Experiments

• First experiment: simple translation

```xml
<vector id="C201" size="17">
  <c id="at_least_one" w="1"/>
  <c id="be" w="1"/>
  <c id="cause" w="2"/>
  <c id="document" w="1"/>
  <c id="domestic" w="1"/>
  <c id="fire" w="3"/>
  <c id="general" w="1"/>
  <c id="home" w="1"/>
  <c id="house" w="1"/>
  <c id="instance" w="1"/>
  <c id="main" w="1"/>
  <c id="mention" w="1"/>
  <c id="private" w="1"/>
  <c id="probable" w="1"/>
  <c id="reference" w="1"/>
  <c id="relevant" w="1"/>
  <c id="specific" w="1"/>
</vector>
```

```xml
<vector id="C201" size="74">
  <!-- Translation of id="fire" w="3" -->
  <c id="allumer" w="3"/>
  <c id="tir" w="3"/>
  <c id="embraser" w="3"/>
  <c id="feu" w="3"/>
  <c id="tirer" w="3"/>
  <c id="incendie" w="3"/>
  <!-- Translation of id="cause" w="2" -->
  <c id="occasionner" w="2"/>
  <c id="provoquer" w="2"/>
  <c id="causer" w="2"/>
  <c id="sujet" w="2"/>
  <c id="procès_" w="2"/>
  <c id="cause" w="2"/>
  <c id="donner" w="2"/>
  ...
</vector>
```
Association Rules: meaning

- **Support**($X \leftrightarrow Y$): the probability $X$ and $Y$ appears together in a transaction.
  - Used to eliminate rare or too frequent occurrences.
  - All supports get lower when the number of transactions raises: in practice we use absolute value in place of ratio.

- **Confidence**($X \Rightarrow Y$): the probability that $Y$ appears knowing that $X$ is in the transaction.
  - A probabilistic dependency from $X$ to $Y$.
  - Less dependent from the number of transactions.
  - High values are preferred.
Association Thesaurus

- Hypothesis: a document is a transaction, set of words forms a consistent set of information
- Production of a graph of terms
  - Link related to “some” semantic, no types
- Using syntactic parsing helps reduction of noise, meaningless relations
- For CLEF: confidence between 20% and 90%
- Possible Use of AT:
  - 2003 Monolingual Query expansion: add related terms
  - 2004 Bilingual Query precision: alignment of two thesaurus, choose the best translation
Multilingual Experiments

• Second experiment: weighted translation
  – Each translation is weighted
  – Using an association thesaurus
  – Idea: w -> t1, ... tn
    • Give a bonus to ti if it has a finite distance with other translations in an association thesaurus.
    • Hypothesis: if 2 words are close in context, their translations are close in context
Association Thesaurus & disambiguation

- Build one Association Thesaurus for each language using all documents
- Hypothesis:
  - the context of a term expresses its semantics
  - each arc of the thesaurus bears one of the meanings of the associated terms
- Thesaurus alignment
  - Associate each couple of term (A,B), in relation in the source thesaurus by a set of couples (X,Y) in the target thesaurus
  - Select (X,Y) with a minimal distance in target thesaurus
  - Meaning: when A is used with B, the X is the best translation of A and Y, of B
Multilingual Experiments

• Example:
  – Find some information about Tamil Tiger suicide bomb attacks or kamikaze actions in Sri Lanka.

<!-- Translation of id="action" w="1" -->
<c id="procès" w="0.166666666666667"/>
<c id="acte" w="0.166666666666667"/>
<c id="empire" w="0.166666666666667"/>
<c id="action" w="0.5"/>
<c id="plainte" w="0.166666666666667"/>
<c id="influence" w="0.166666666666667"/>
Multilingual Experiments

• But:
  – Results got worse!

• Because:
  – Quality of the dictionaries
  – Quality/Size of the thesaurii
    • Too few entries in the thesaurus (~4000 to ~9000)
  – Most of the time, selected translations are the most frequent translations but selection does not really depend on the context...

• However
  – Trowing out the thesaurii to directly take into account the context of translations may still be a good idea.
Results

• Drop mono-> bilingual
  – (Eng) Russian: 35% -> 11%, 4% with thesaurus..
  – (Fr) Russian: 35% -> 6%, 5% with thesaurus

• Possible explanation
  – Division by the number of translation reduce importance of possible tool words
  – Raising weight of correct translation works on terms with many translation hence give more importance to words not really topic related
Conclusion

- Correct results on monolingual track
  - Effectiveness of syntactic parsing + DFR
- Bad results on bilingual track
  - Manual checking are good …
  - Not due to weighting (cf. monolingual)
  - Possible wrong re-weighting method
  - Not enough linguistic resources?
  - Possible experimentation error
  - Wrong hypothesis on only one sense in corpus
What next?

• Redo the experiment of parallel bilingual thesaurus
  – Understand what is wrong
  – Have better linguistic resources (but how?)

• Better use of the output of the parser
  – Using noun phrase to enhance precision