Using Statistical Translation Models for Bilingual IR

Jian-Yun Nie, Michel Simard

Lab. RALI
Département d'Informatique et Recherche opérationnelle,
Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal, Québec, H3C 3J7 Canada
{nie, simardm}@iro.umontreal.ca

Context and Goals

Context: We developed an automatic mining system for parallel texts on the Web - PTMiner.

Goal: Further test how effective a mined parallel corpus and the resulting statistical translation model are for CLIR.

Tests:

- cleaning of parallel corpora
- cutoff translation models
- two-directional query translation
- combination of translation models with dictionaries

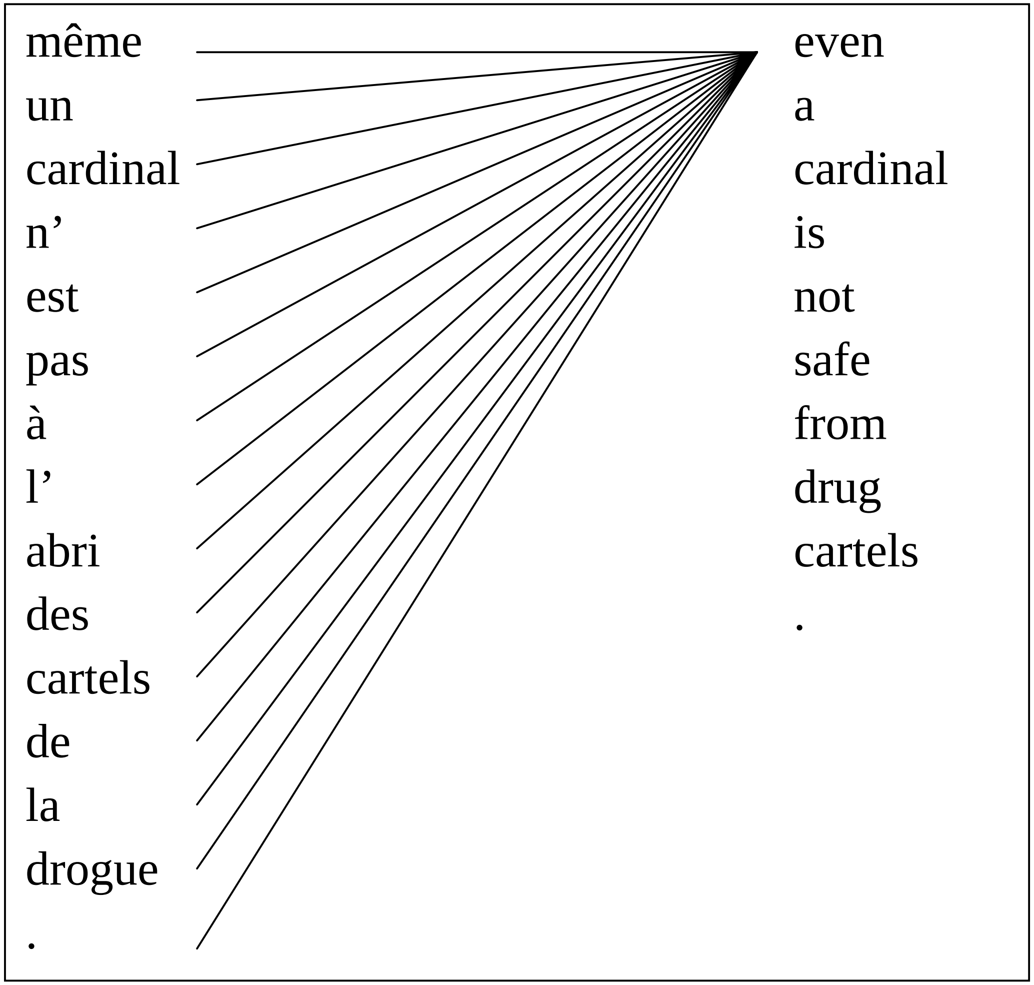
A quick view on PTMiner

- Determination of potential web sites for parallel web pages
- Crawling the candidate sites
- Examination of parallelism
  - length
  - HTML markers
  - (sentence alignment)
- Precision estimated at 80%

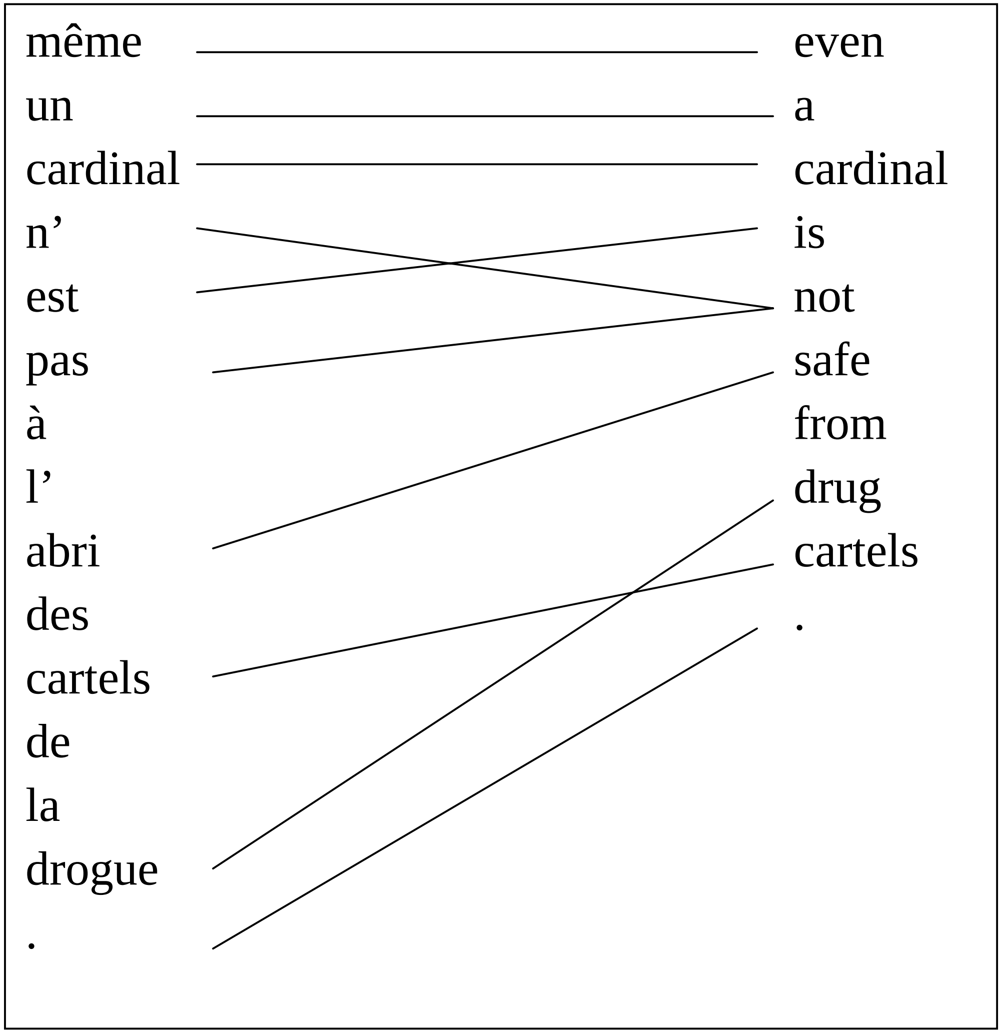
Model training

- $p(e_j|f_i)$ is estimated from a parallel training corpus, aligned into parallel sentences
- No syntactic features and position information (IBM model 1)
- Process:
  - Input = two sets of parallel texts
  - Sentence alignment A: $\mathrm{E_k} \leftrightarrow \mathrm{F_l}$
  - Initial probability assignment: $t(e_j|f_i, A)$
  - Expectation Maximization (EM): $p(e_j|f_i , A)$
  - Final result: $p(e_j|f_i) = p(e_j|f_i , A)$

Initial probability assignment $t(e_j|f_i, A)$



Application of EM: $p(e_j|f_i, A)$



même
un
cardinal
n’
est
pas
à
l’
abri
des
cartels
de
la
drogue
.

even
a
cardinal
is
not
safe
from
drug
cartels
.

Size of the corpora

	E-F		E-G		E-I	
Text Pairs	18 807		10 200		8 504	
Size (Mb)	174	198	77	100	50	68

Model cutoff

- Observation: Low probability translations are often bad translations.
- Size constraints in practical uses.
- Filter out bad translations by
  - eliminating low probability translations (threshold)
  - Fix the size of the model and eliminate the entries that impact the model the least.

Results on CLEF2000 with cutoffs

	1M	100K	10K	5K	1K	P≥0.05	P≥0.1	P≥0.25
de-en	0.1684	0.1559	0.1403	0.1212	0.0714	0.1693	0.1651	0.1640
it-en	0.2442	0.2237	0.2426	0.2059	0.0989	0.2444	0.2524	0.2393
fr-en								

Corpus cleaning

- About 20% of the original corpus is noise
- Eliminate the noisy part of the corpus by:
  - trying to align sentences (length-based alignment)
  - considering “known translations” (increase alignment score)
- If unaligned sentences in a text pair larger than a threshold, then remove the pair.

Experiments on Chinese-English

Direction	No filter	Best filtering
E-C	161 (80.50%)	183 (91.50%)
C-E	154 (77.00%)	173 (86.50%)

Translation accuracy of first translations
of 200 random words

C-E CLIR results

Direction	No filter	Best filtering
E-C	0.1843 (47.11%)	0.2013 (50.63%)
C-E	0.1898 (49.16%)	0.2063 (53.43%)

Some improvements after cleaning

CLEF 2000 after cleaning

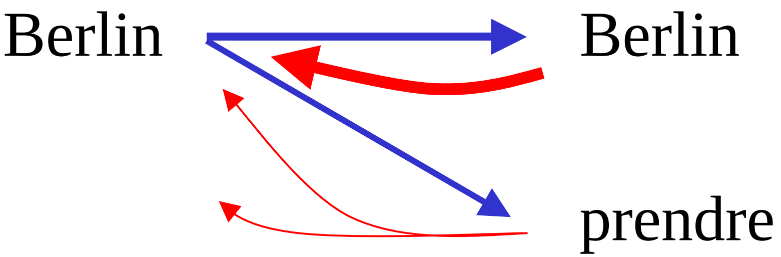
	1M	100K	P≥0.05	P≥0.1	P≥0.25
de-en	0.0764	0.0745	0.0777	0.0751	0.0669
it-en	0.2209	0.2418	0.2453	0.2448	0.2363
fr-en					

Degradation of performance,

in particular for de-en

Two-directional translation

- Some common words often appear as top translations (e.g. prendre) because they often co-occur in parallel corpora with many source words.
- However, their translation back to the source language will be sparse.
- Considering the backward translation may eliminate such words and return stronger 1 - 1 translations.



Results with two-directional translation

	1M	100K	10K	5K	1K	P≥0.05	P≥0.1
de-en	0.1026	0.1337	0.1339	0.1138	0.0545	0.1259	0.1257
it-en	0.2116	0.2149	0.2182	0.1971	0.0945	0.2185	0.2181
fr-en							

Degradation w.r.t. one-directional translation

Submitted runs

- 3 sets of bilingual runs fr-en, de-en and it-en
  - Translation with model P≥0.1
  - Combination with dictionaries (FreeDict) and assign every dictionary translation with equal weight (0.001)
  - Combination with dictionaries and assign the weight of *idf* to every dictionary translation

Average precision of the submissions

	RaliP01	RaliM001	RaliMidf
fr-en	0.3499	0.3564	**0.3685**
de-en	0.2124	0.2188	**0.2565**
it-en	0.2731	**0.2742**	0.2562

Comparison with medium run

	RaliMidfF2E	RaliMidfD2E	RaliM001D2E
≥ medium	41	27	27
< medium	6	20	20

Trans. From Italian: Mad cow desease in Europe

europe=0.382011
europa=0.107791
pazzi=0.083633
vaild=0.080209
bunch=0.080209
lot=0.077385
cow=0.066805
chance=0.064079
paziente=0.057877
europe=0.133206
find=0.128462
case=0.109291
document=0.089954
acknowledgement=0.077600
documentation=0.038357

Trans from French: IRA attack of airport

airport=0.593288
attack=0.240423
bomb=0.092175
people=0.074114
airport=0.203591
europe=0.177602
describe=0.148660
act=0.134723
commit=0.123677
find=0.122739
terrorism=0.065951
european=0.023055

Observations

- Translation models seem to work well for en-fr (better than en-de and en-it).
  - Corpus size is not a factor.
  - Corpus quality?
  - We have good morphological transformer for English and French.
- Simple stemmers are used for German and Italian.
  - Problematic for German:

    elektroschwachtheorie, kriegsdienstverweigerer, welthandelsorganisation, ...

Observations (cont'd)

- Corpus cleaning did not help. (Any error or new parameters?)
- Two-directional query translation did not work well. (Any error?)
- Model cutoffs improve CLIR effectiveness, in particular by a probability threshold.
- Future work:
  - Translation models integrating compound terms may bring some further improvement.
  - Translation filtering
  - Mining larger corpora and for more languages
  - Better integration with dictionaries