Automatic Identification of Cognates and False Friends in French and English

Diana Inkpen and Oana Frunza
University of Ottawa
and
Greg Kondrak
University of Alberta

Second-language learning

- When learning a second language a student can benefit from knowledge in his/her first language (Gass,1987), Ringbom,1987).
- The use of cognates in second language teaching was shown to accelerate vocabulary acquisition and to facilitate reading comprehension (LeBlanc et al., 1989).

- Morphological rules for conversion from English to French also proved to help (Treville, 1990).
- Manual work on cognate detection: LeBlanc and Seguin (1996) collected 23,160 French-English cognate pairs from two general-purpose dictionaries (70,000 entries). Cognates make up over 30% of the vocabulary.

French and English

- Although French and English belong to different branches of the Indoeuropean family of languages, they share a very high number of cognates.
- The majority are words of Latin and Greek origin: éducation - education and théorie - theory.
- A small number (120) of very old, "genetic" cognates go back all the way to Proto-Indoeuropean, mère - mother and pied - foot.
- Other cognates can be traced to the conquest of Gaul by Germanic tribes after the collapse of the Roman Empire, and by the period of French domination of England after the Norman conquest.

Definitions

- *Cognates* or *True Friends* (*Vrais Amis*), words that are similar and are mutual translations
  nature - nature, recognition - reconnaissance
- *False Friends* (*Faux Amis*) words that are similar but have different meanings
  main "hand" - main, blesser "to injure" - bless (bénir)
- *Partial Cognates* words that have the same meaning in both languages in some but not all contexts
  facteur - factor or "mailman"
- *Genetic Cognates* - derived directly from the same word in the ancestor language (may differ in form or meaning)
  père - father, chef - head

Related work
Bilingual corpora and translation lexicons

- Simard et al, (1992) use cognates to align sentences in bitexts (cognates = first 4 characters are identical).
- Brew and McKelvie (1996) extract French-English cognates and false friends from bitexts using a variety of orthographic similarity measures.
- Mann and Yarowsky (2001) automatically induce translation lexicons on the basis of cognate pairs.
- Kondrak (2004) identifies genetic cognates in vocabularies of related languages by combining the phonetic similarity of lexemes with the semantic similarity of glosses.

Our task

- **Automatic identification of cognates and false friends** => lists, for inclusions in learning aids
- For a pair of word, the two classes for the automatic classification are: **Cognates/False-Friends** and **Unrelated**.
- Additional "translation" feature:
  - Cognates if the two words are translations of each other in a bilingual dictionary;
  - False Friends otherwise.

Our method

- Machine learning (Weka)
- Features: 13 orthographic similarity measures.
  - Each measure separately.
  - Average.
  - Combine them through ML algorithms.

Orthographic similarity measures

- IDENT returns 1 for identical words, 0 otherwise.
- PREFIX returns the length of the common prefix divided by the length of the longer string.
- DICE divides twice the number of shared letter bigrams by the total number of bigrams in both words.
  DICE(*colour*, *couleur*) = 6/11 = 0.55
  (the shared bigrams are *co*, *ou*, *ur*)
- TRIGRAM same as DICE but uses trigrams.
- XDICE uses trigrams without the middle letter.
- XXDICE takes into account the positions of bigrams.

Orthographic similarity measures (cont.)

- Longest Common Subsequence Ratio (LCSR)
  LCSR(*colour*, *couleur*) = 5/7 = 0.71
- Normalized Edit Distance (NED)
- SOUNDEX – phonetic name matching, numeric codes.
- BI-SIM, TRI-SIM, BI-DIST, and TRI-DIST generalize LCSR and NED measures, uses letter bigrams or trigrams instead of single letters.

Training data

- http://mypage.bluewin.ch/a-z/cusipage/basicfrench.html bilingual list of 1047 basic words and expressions. We excluded multi-word expressions. We manually classified 203 pairs as Cognates and 527 pairs as Unrelated.
- A manually word-aligned bitext (Melamed, 1998).We manually identified 258 Cognate pairs.
- A set of exercises for Anglophone learners of French (Treville, 1990) (152 Cognate pairs).
- An on-line (http://french.about.com/library/fauxamis/blfauxam.htm) list of "French-English False Cognates" (314 False-Friends).

Test data

A separate test set extracted from the following sources:

- A random sample of 1000 word pairs from an automatically generated translation lexicon.
- We manually classified 603 pairs as Cognates and 343 pairs as Unrelated.
- The on-line list of “French-English False Cognates” (94 additional False-Friends).

Data (summary)

	Training set	Test set
Cognates	613 (73)	603 (178)
False-Friends	314 (135)	94 (46)
Unrelated	527 (0)	343 (0)
Total	1454	1040

Results of classification

Orthographic similarity measure	Threshold	Accuracy on Training set	Accuracy on Test set
IDENT	1	43.90 %	55.00 %
PREFIX	0.03845	92.70 %	90.97 %
DICE	0.29669	89.40 %	93.37 %
LCSR	0.45800	92.91 %	94.24 %
NED	0.34845	93.39 %	93.57 %
SOUNDEX	0.62500	85.28 %	84.54 %
TRI	0.0476	88.30 %	92.13 %
XDICE	0.21825	92.84 %	94.52 %
XXDICE	0.12915	91.74 %	95.39 %
TRI-SIM	0.34845	95.66 %	93.28 %
TRI-DIST	0.34845	95.11 %	93.85 %
Average measure	0.14770	93.83 %	94.14 %

Results of classification

Classifier	Accuracy cross-val. on training set	Accuracy on test set
Baseline	63.75 %	66.98 %
OneRule	95.66 %	92.89 %
Naive Bayes	94.84 %	94.62 %
Decision Tree	95.66 %	92.08 %
Decision Tree (pruned)	95.66%	93.18 %
IBK	93.81 %	92.80 %
Ada Boost	95.66 %	93.47 %
Perceptron	95.11 %	91.55 %
SVM (SMO)	95.46 %	93.76 %

Pruned Decision Tree

```
TRI-SIM <= 0.3333
|   TRI-SIM <= 0.2083: UNREL (447.0/17.0)
|   TRI-SIM > 0.2083
|   |   XDICE <= 0.2: UNREL (97.0/20.0)
|   |   XDICE > 0.2
|   |   |   BI-SIM <= 0.3: UNREL (3.0)
|   |   |   BI-SIM > 0.3: CG_FF (9.0)
TRI-SIM > 0.3333: CG_FF (898.0/17.0)
```

Results on genetic cognates set

Classifier	Accuracy
Baseline	—
OneRule	35.39 %
Naive Bayes	29.20 %
Decision Tree	35.39 %
Decision Tree (pruned)	38.05 %
IBK	43.36 %
Ada Boost	35.39 %
Perceptron	42.47 %
SVM (SMO)	35.39 %

120 pairs of genetic cognates, available at:

http://www.cs.ualberta.ca/

~kondrak/cognatesEF.html

Conclusion

- We presented several methods to automatically identify cognates and false friends.
- We tested a number of orthographic similarity measures individually. We combined them using several different machine learning classifiers.
- We evaluated the methods on a training set, on a test set, and on a list of genetic cognates.
- The results show that, for French and English, it is possible to achieve very good accuracy even without the training data by employing orthographic measures of word similarity.

Future work

- Automatically identify partial cognates (WSD).
- We plan to use translation probabilities from a word-aligned parallel corpus.
- Produce complete lists of cognates and false friends, given two vocabulary lists for the two languages.
- Apply our method to other pairs of languages (since the orthographic similarity measures are not language-dependent).
- Include our lists of cognates and false friends into language learning tools.