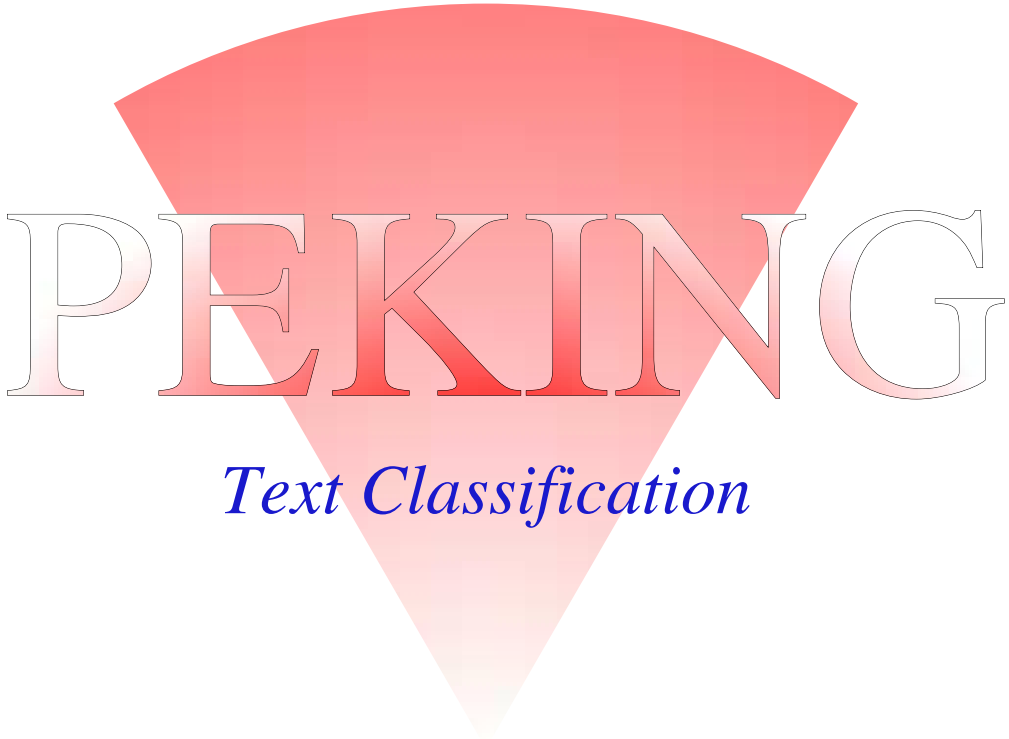
Cross-Lingual Text Categorization

Nuria Bel and Marta Villegas, gilcUB Barcelona
Cornelis H.A. Koster, University of Nijmegen



overview

- Cross-lingual Text Categorization
- The experimental procedure
- The mono-lingual baseline
- Poly-lingual training and testing
- Conclusions

The problem

An organization, which already has an automatic classification system installed, wishes to extend this system to classify also documents in other languages.

In order to ease the transition, some documents in those other languages are provided, either in untranslated form but manually supplied with a class label, or in translated form and without such a label.

With minimal manual intervention, a bootstrap of the system must be performed, so that documents in all those languages can be classified automatically in their original form by a single poly-lingual classifier.

Cross-lingual Text Categorization?

- Cross-Lingual Information Retrieval (CLIR)

  a user formulates a query in one language in order to retrieve documents in several (other) languages

Cross-lingual Text Categorization?

- Cross-Lingual Information Retrieval (CLIR)
- Cross-Lingual Text Categorization

a classifier is trained to classify documents in several languages

Cross-lingual Text Categorization?

- Cross-Lingual Information Retrieval (CLIR)
- Cross-Lingual Text Categorization
- Differences and agreements?

both are based on a computation of similarity
CLIR is based on queries consisting of a few words
in CLTC each class is defined by an extensive profile

Cross-lingual Text Categorization?

- Cross-Lingual Information Retrieval (CLIR)
- Cross-Lingual Text Categorization
- Differences and agreements?
- Can CLTC learn from CLIR?

need similar linguistic resources
unlike CLIR, CLTC needs no feedback techniques
in CLTC the profiles act like a language model

The classification experiment

- the ILO corpus

  2165 english documents and 1590 spanish documents
  monoclassified into 12 classes of varying size
  around 2000 words per document

The classification experiment

- the ILO corpus
- the Linguistic Classification System LCS

Winnow and Rocchio algorithms
using optimal tuning

The classification experiment

- the ILO corpus
- the Linguistic Classification System LCS
- experimental procedure

  training in 25/75 or 50/50 shuffles
  using heavy cross-validation
  allowing 0-3 classes per document
  accuracy: micro-averaged F1-value

ILO overview

class name	# docs English	# docs Spanish	class description
02	123	74	Human rights
03	397	86	Conditions of employment
04	299	71	Conditions of work
05	22	23	Economic and social development
06	414	448	Employment
07	279	278	Labour Relations
08	85	81	Labour Administration
09	98	86	Health and Labour
10	156	148	Social Security
11	81	20	Training
12	131	154	Special prov. by category of persons
13	108	121	Special prov. by Sector of Econ. Act.
Total:	2165	1590	

the mono-lingual baseline

comparing

- English and Spanish
- Winnow and Rocchio
- three different document representations.

Three document representations

- keywords

  all words of the text

  minimal preprocessing, no lemmatization

Three document representations

- keywords
- lemmatized keywords

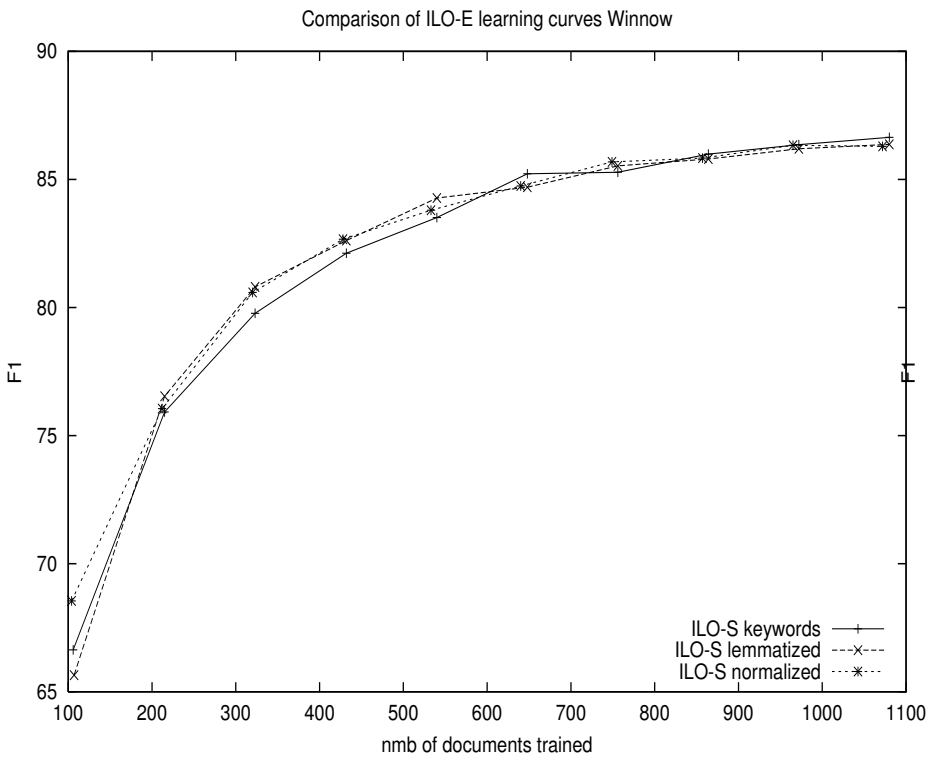
same preprocessing plus lemmatization

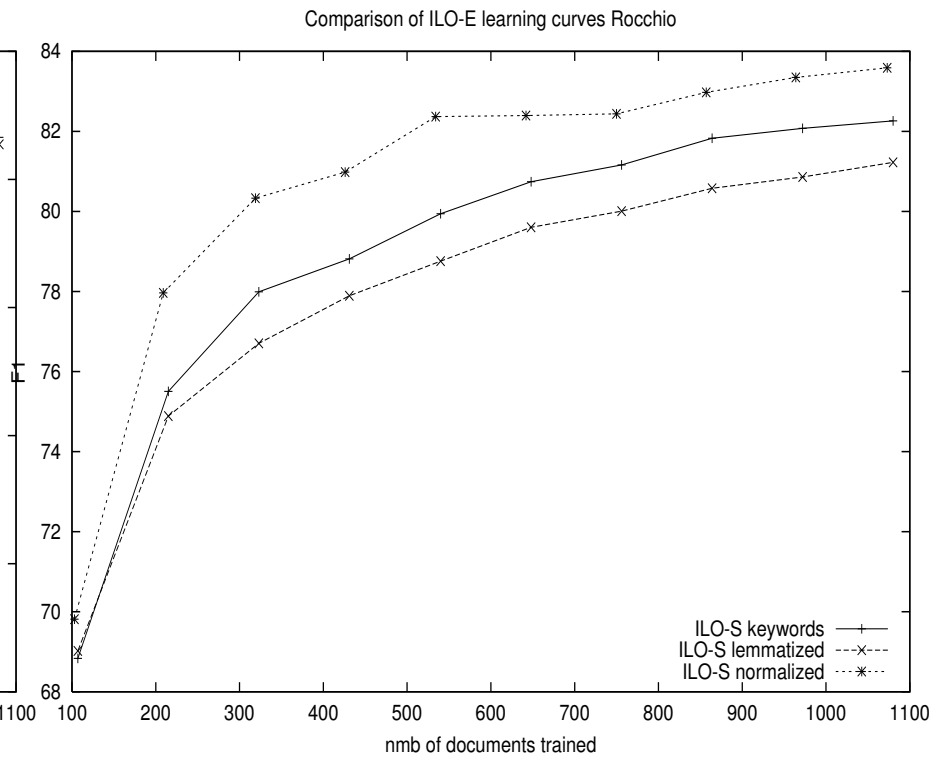
Three document representations

- keywords
- lemmatized keywords
- linguistically motivated terms

  combining mutiword phrases into one term ("normalization")
  like ’software_engineering’ or ’Trabajadores_migrantes’
  linguistic resources developed by gilcUB

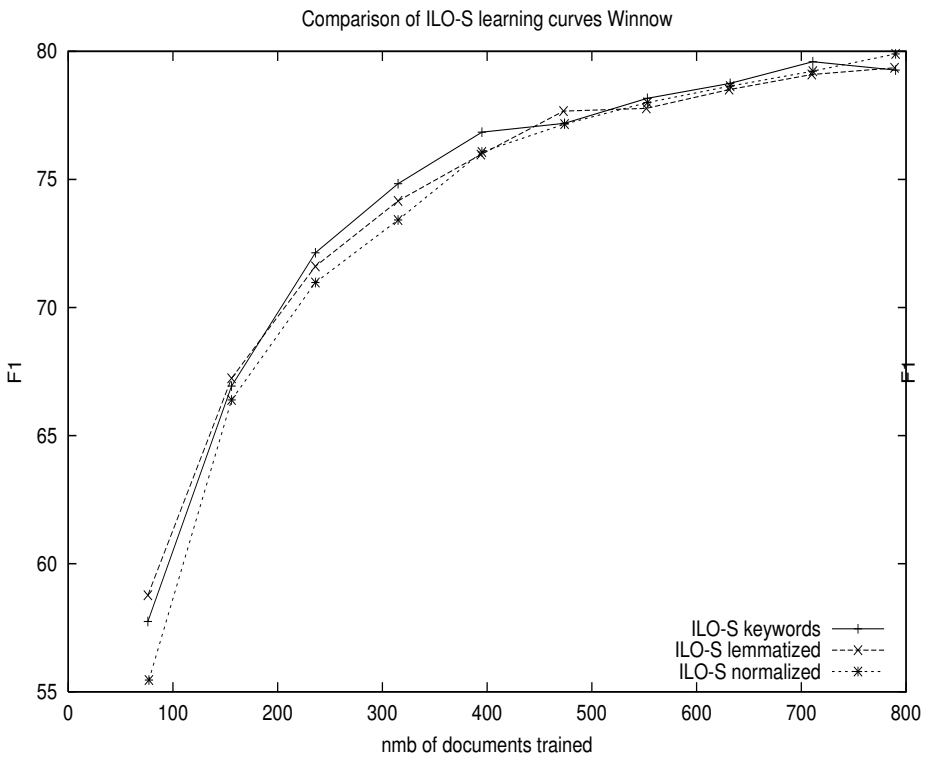
learning curves (English)

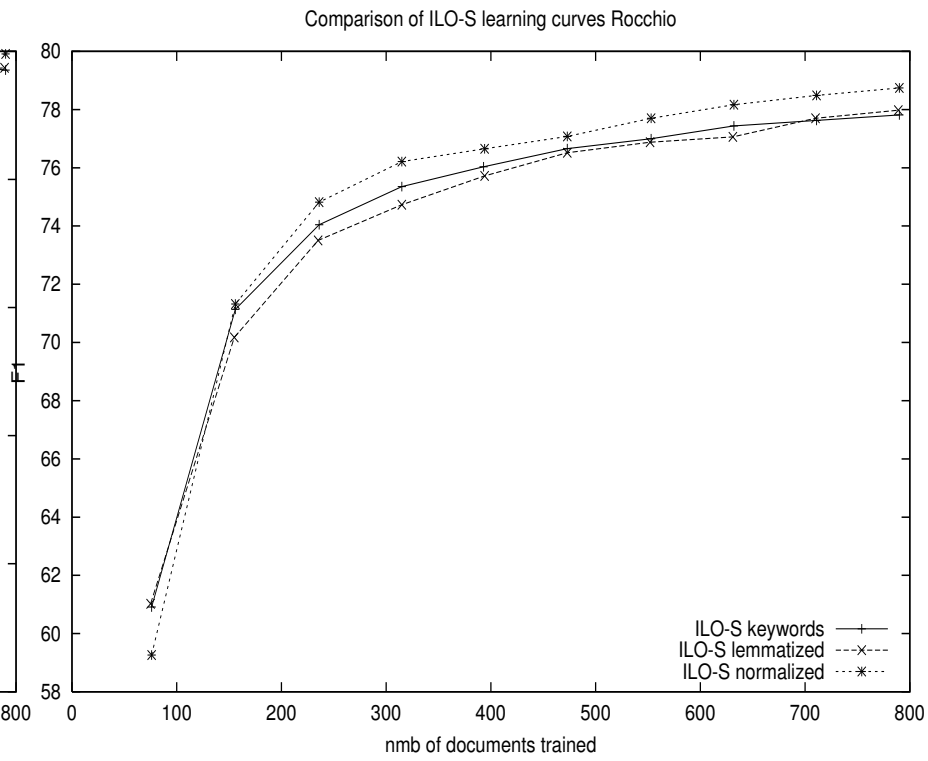




- for Winnow representation makes no difference
- for Rocchio some (just significant) effects

learning curves (Spanish)

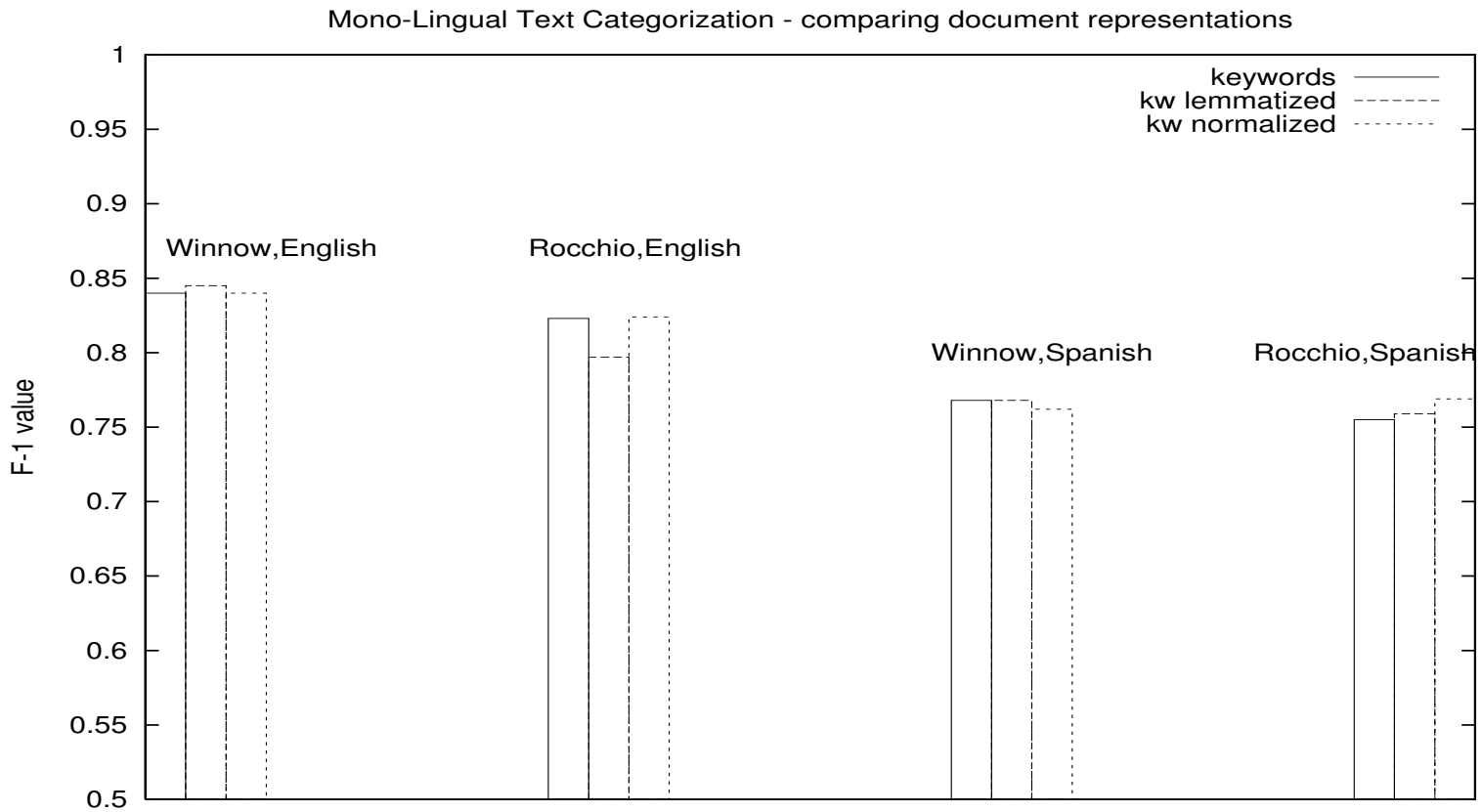




- worse than English, but fewer documents
- effect of linguistic improvements is disappointing.

Comparing document representations

- training on 25% of the documents, testing on the other documents, 12-fold cross-validation



Cross-lingual approach

translation strategies:

- **no translation**

  needs enough labeled training documents in new language

  poly-lingual training

Cross-lingual approach

translation strategies:

- **no translation**
- **document translation**

automatic translations not very satisfactory
manual translations are too expensive
we did not try this approach.

Cross-lingual approach

translation strategies:

- **no translation**
- **document translation**
- **terminology translation**

  constructing a terminology for each of the categories
  translating all domain terms
  will include all or most of the terms which are relevant for classification.

Cross-lingual approach

translation strategies:

- **no translation**
- **document translation**
- **terminology translation**
- **profile-based translation**

  translate only the terms actually occurring in the class profiles
  the most important 150 terms per class, 923 different terms

The linguistic resources

Terminology translation Spanish-English and vv

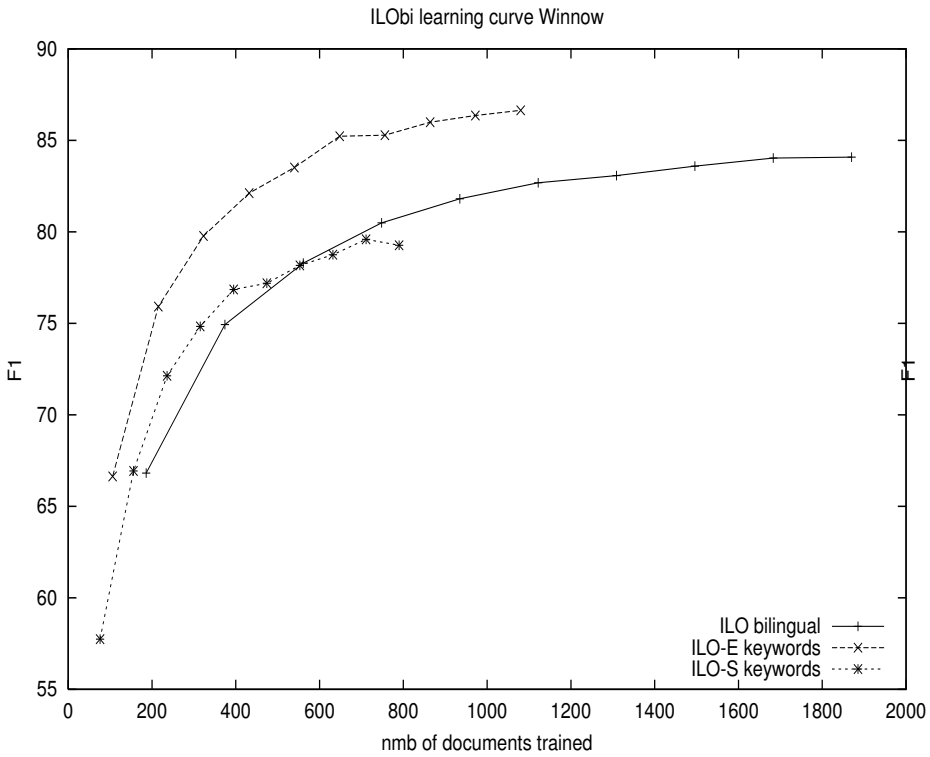
- corpus-driven approach
- nouns, adjectives and verbs with frequency higher than 30
- 4462 wordforms (out of 4.619.681 tokens) for Spanish
- 5258 (out of 4.609.670 tokens) for English.

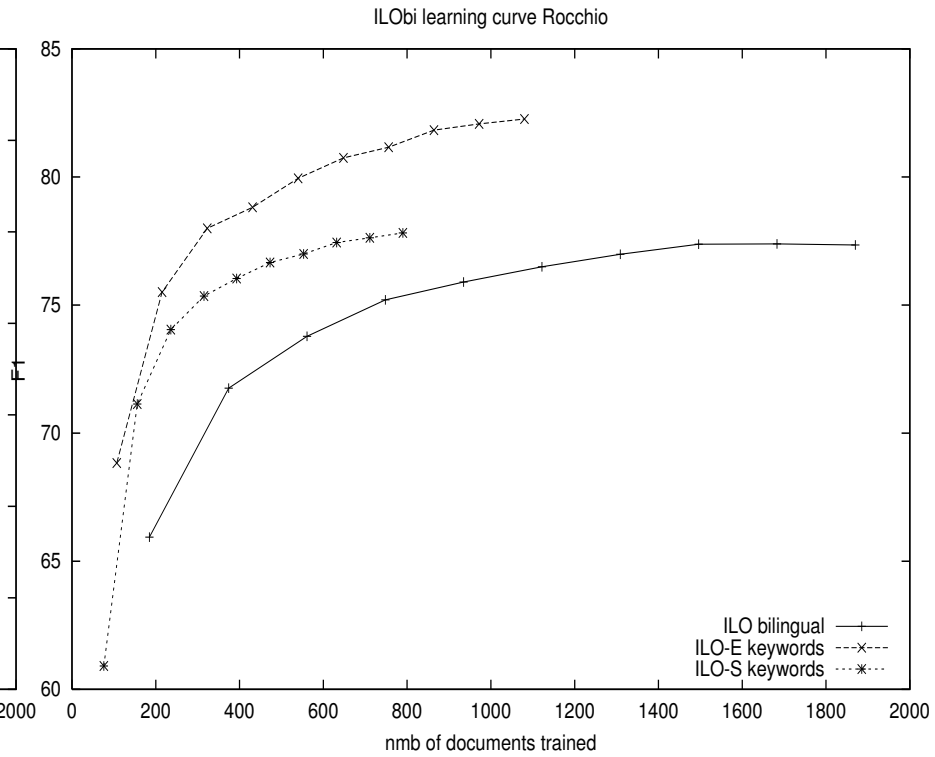
Poly-lingual training (1)

- building a single classifier
- from labeled train documents in a mix of languages
- which will classify documents in any of the trained languages
- without translating anything
- even without trying to find out what language the documents are in
- using no linguistic resources.

Poly-lingual training (2)

- mixing 2167 English and 1590 Spanish ILO documents





- Winnow copes well, Rocchio fails.

Terminology translation (1)

1. training a classifier on all 2167 normalized English documents
2. using this classifier to classify the 1590 pseudo-English (Spanish) documents.

algorithm	representation	language	Accuracy	
			Multi 0:3	Mono 1:1
Winnow	keywords	English and pseudo-English	.696±.051	.792±.012
Rocchio	keywords	English and pseudo-English	.592±.025	.709±.012
Winnow	keywords	Spanish and pseudo-Spanish	.552±.062	.617±.062
Rocchio	keywords	Spanish and pseudo-Spanish	.538±.045	.589±.029

Terminology translation (2)

- Winnow much better than Rocchio
- mono-classification much better than multi
  - thresholds wrong due to untranslated words
  - word distribution in train- and test set different
- terminology translation is a viable approach for cross-lingual mono-classification.

Profile-based translation (1)

1. we took the best 150 terms from each class profile and combined the results into a vocabulary of 923 different words (out of 22000)
2. some of them quite surprizing (a1 next)
3. a classifier was trained on all English documents but using only the words in the vocabulary
4. this classifier was tested on all Spanish documents, translating only the Spanish terms having a translation towards a word in the vocabulary.

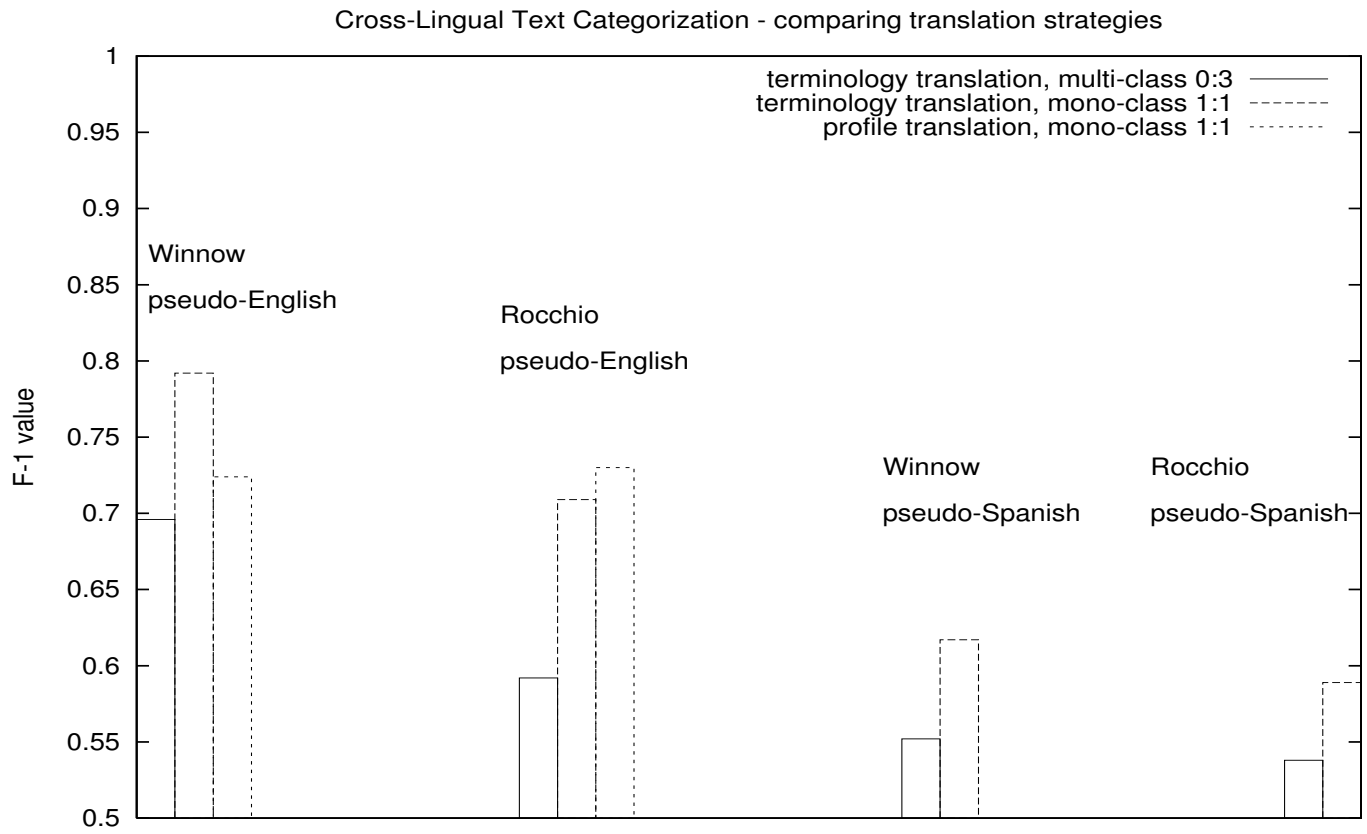
Profile-based translation (2)

- training on English documents and classifying Spanish documents in which just the profile words have been translated
- only a small translation effort needed.

algorithm	representation	language	Accuracy	
			Multi 0:3	Mono 1:1
Winnow	keywords	Eng/Spa	.605±.071	.724±.035
Rocchio	keywords	Eng/Spa	.681±.048	.730±.019

Comparing translation results

- training on 25% of the documents in one language, testing on all documents in the other, 12-fold cross-validation



Conclusions

- CLTC is easier than CLIR

  the law of large numbers is with us

Conclusions

- CLTC is easier than CLIR
- no need for conflation of synonyms or lemmatization

  if two equivalent forms of a word occur frequently enough to have an impact on classification, they will also do so as independent terms

Conclusions

- CLTC is easier than CLIR
- no need for conflation of synonyms or lemmatization
- poly-lingual training

  training one single classifier to classify documents in a number of languages
  works well for Winnow, not for Rocchio

Conclusions

- CLTC is easier than CLIR
- no need for conflation of synonyms or lemmatization
- poly-lingual training
- terminology translation

  translating just the typical terms of the documents works well

Conclusions

- CLTC is easier than CLIR
- no need for conflation of synonyms or lemmatization
- poly-lingual training
- terminology translation
- profile-based translation

  translating only the terms occurring in the class profile
  less accurate but cheap and simple

Conclusions

- CLTC is easier than CLIR
- no need for conflation of synonyms or lemmatization
- poly-lingual training
- terminology translation
- profile-based translation
- the above techniques can be combined

using terminology translation or profile-based translation to generate examples for poly-lingual training
then bootstrap the poly-lingual classifier
with some manual checking of uncertain classifications.

Conclusions

- CLTC is easier than CLIR
- no need for conflation of synonyms or lemmatization
- poly-lingual training
- terminology translation
- profile-based translation
- the above techniques can be combined

- Cross-Language Text Categorisation is a solved problem!