Label Propagation for Fine-Grained Cross-Lingual Genre Classification

P. Petrenz B. Webber

THE UNIVERSITY of EDINBURGH

xLiTe Workshop on Cross-Lingual Technologies

07 December 2012

- **Genre classification can benefit web search**
  Allows users to filter documents by genre

- **… and inform NLP applications**
  Summarization, Word-Sense Disambiguation, Tagging, etc.

- **Training data available for few languages**
- **Cross-Lingual techniques**
  Exploit labels in another language to predict genres
- **Cross-Lingual Genre Classification (CLGC)**
  Possible without machine translation (Petrenz 2012)
  Work focused on broad categories (2-4 classes)

- **Genres are multi-dimensional**
  Defined by communicative purpose, medium, target audience, topic (?) etc.

- **Idea: Use separate feature sets**
  Different feature types used in mono-lingual classification: Structural, lexical, presentational, etc.

- **Label Propagation**
  Initially propsed by Xhu & Ghahramani (2002)
  Exploits target language texts
  Easily adapted to more than one feature space

Features

Source and target languages

Target language only

Cross-Lingual

Simple statistics (Mean word length, Type/Token ratio etc.)

Frequencies of 12 universal PoS tags (Petrov et al. 2011)

Standardized to remove language bias

PoS histogram

Mean and SD for each PoS tag over sliding windows of 5 tags (Feldman et al. 2009)

Structural feature set

Bag of Words

Binary representation of word occurance

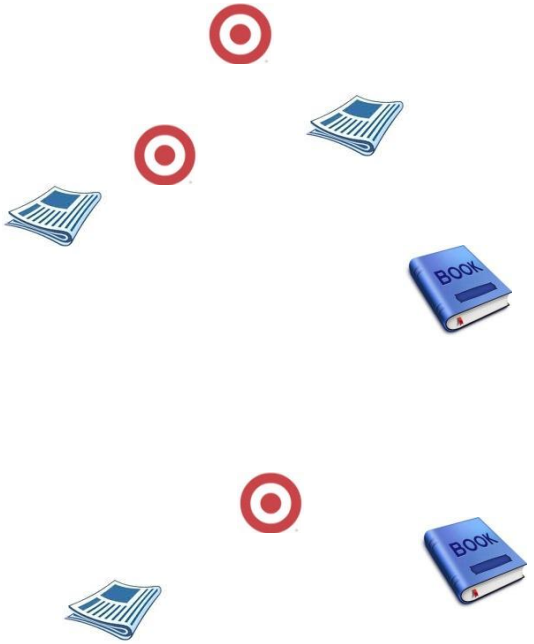
Lexical feature set

Cross-Lingual

PoS histogram

Bag of Words

Label propagation

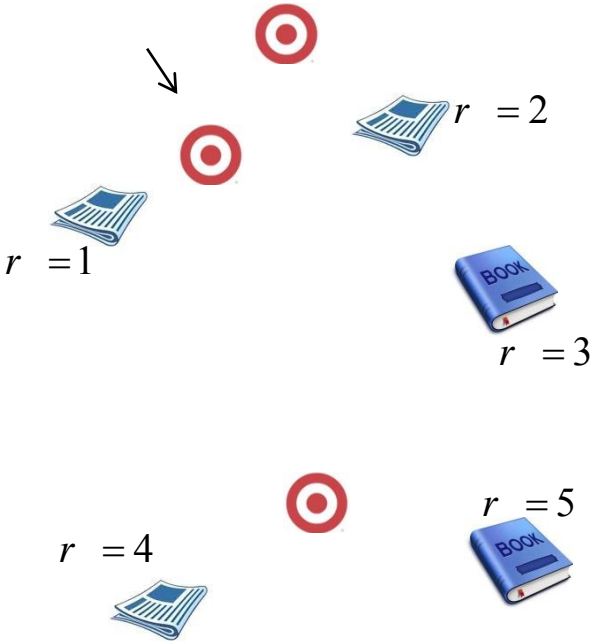


Cross-Lingual

PoS histogram

Bag of Words

Label propagation



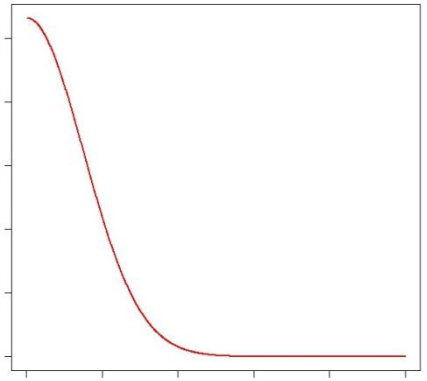
Cross-Lingual

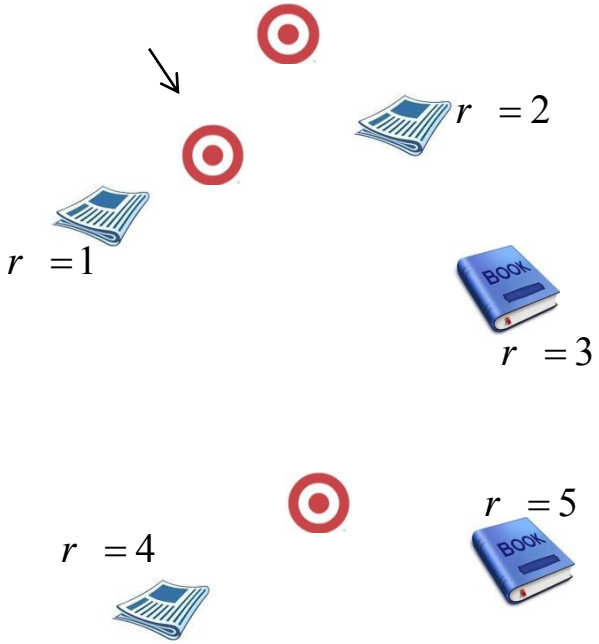
PoS histogram

Bag of Words

Label propagation

$$w_{ij}^{f} = \exp\left(-\frac{(r_{ij}^{f}-1)^2}{2\sigma^2}\right)$$



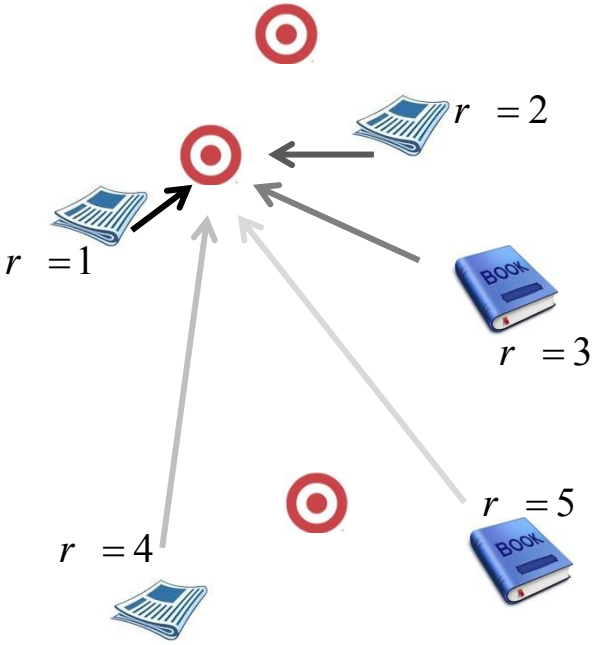


Cross-Lingual

PoS histogram

Bag of Words

Label propagation

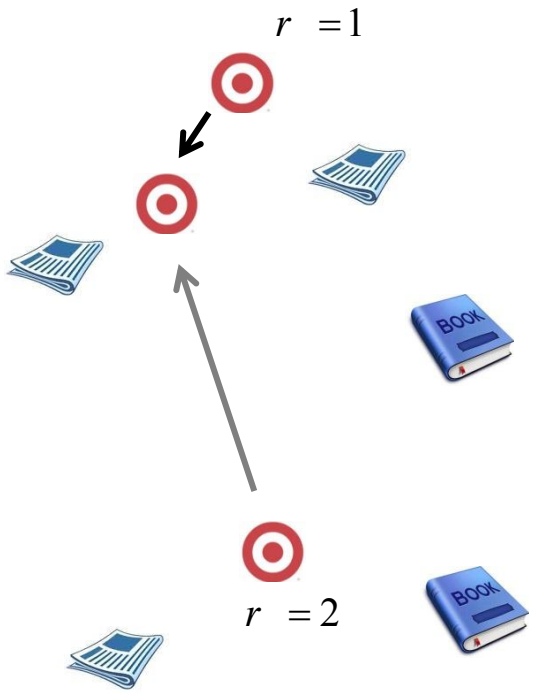


Cross-Lingual

PoS histogram

Bag of Words

Label propagation

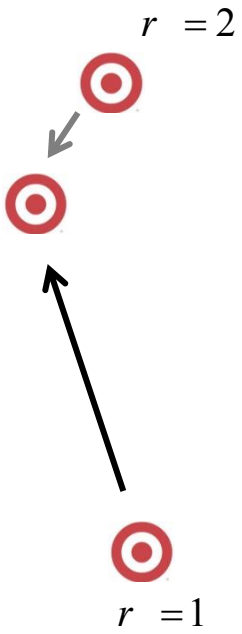


Cross-Lingual

PoS histogram

Bag of Words

Label propagation



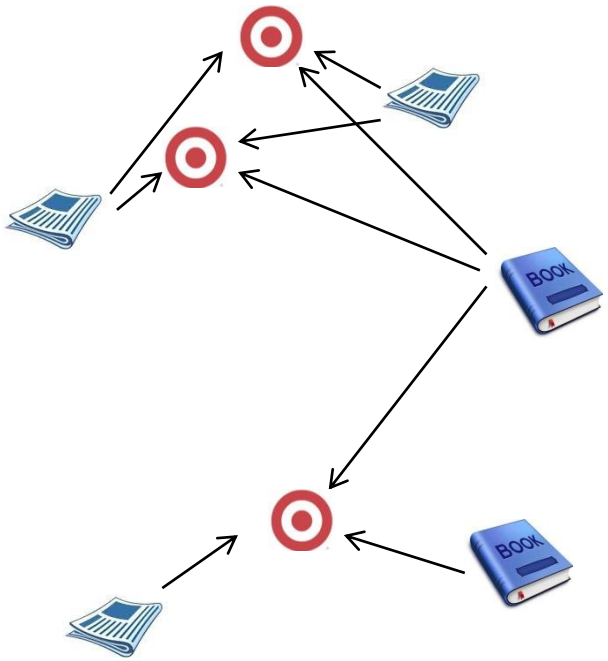
Label propagation

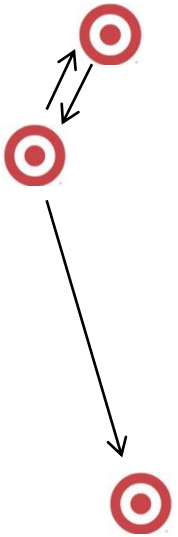
Cross-Lingual

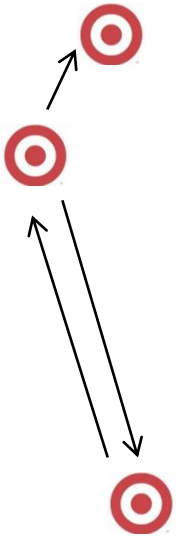
Cross-Lingual

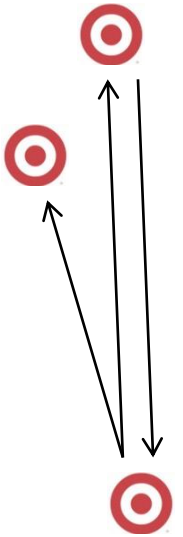
PoS histogram

Bag of Words



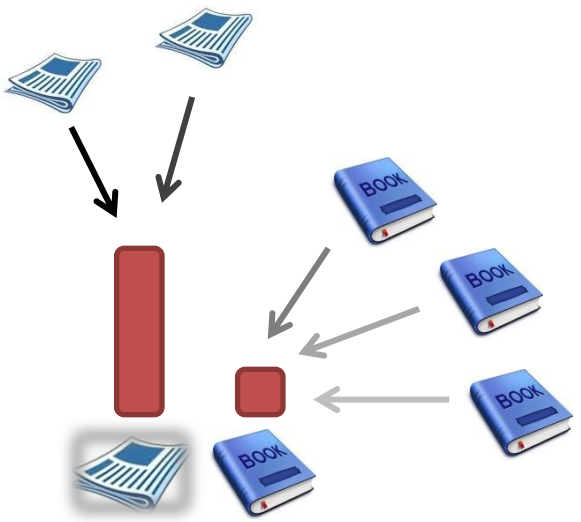






Prior

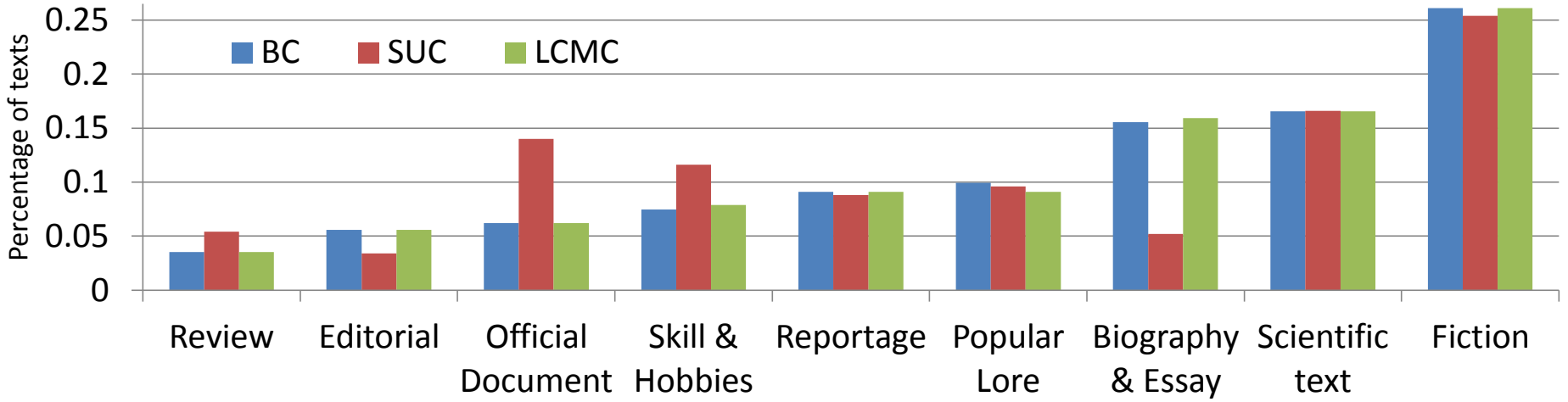
Label propagation



Experiments

English, Swedish, and Chinese texts

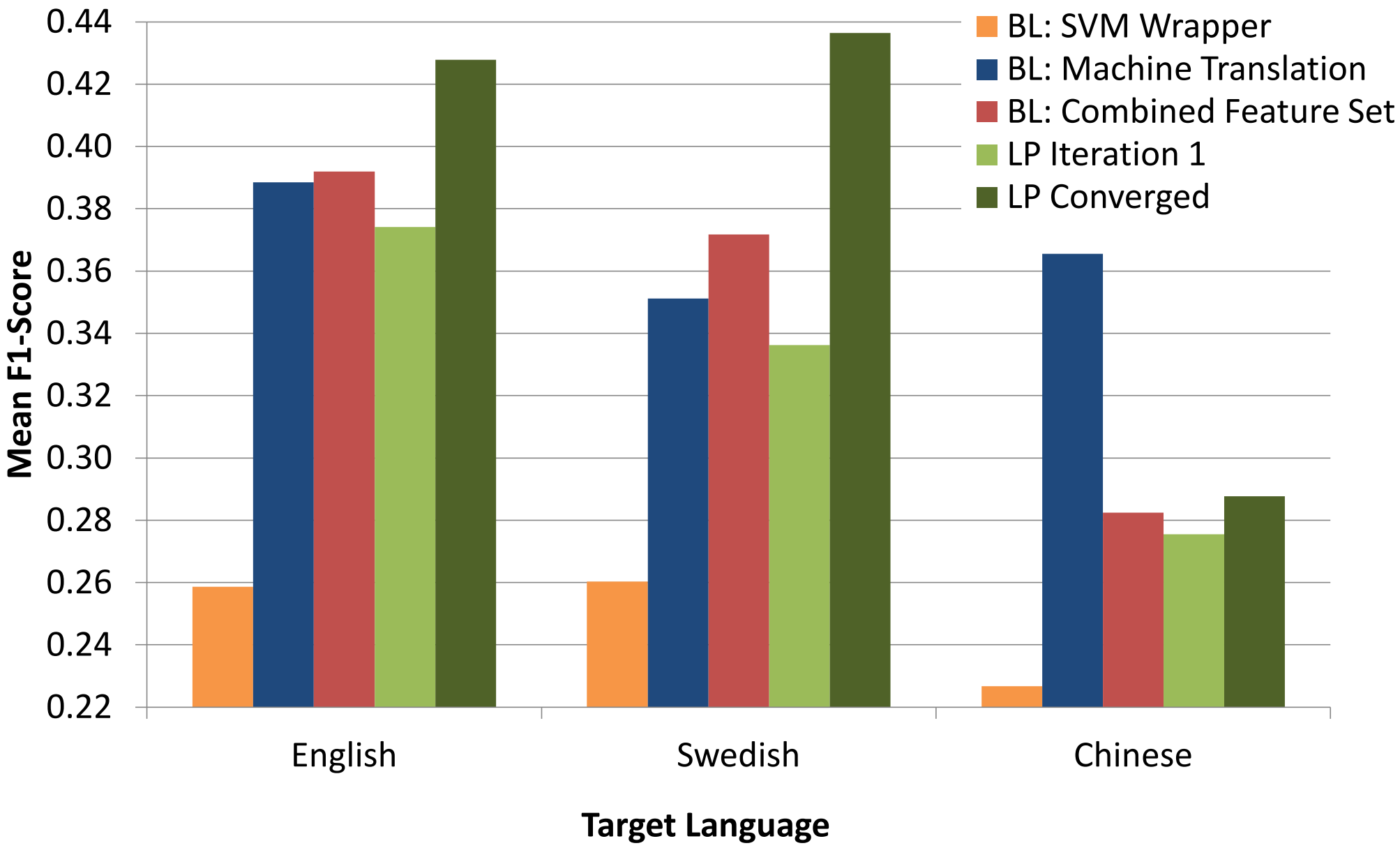
Sources: BC, SUC, LCMC; 9 genre classes; PoS tagged



Baselines

1. SVM wrapper algorithm (Petrenz 2012)
2. Full text machine translation + mono-lingual classifier
3. Label Propagation with combined feature set

Experiments



Conclusion

- **Fine-grained CLGC possible**

  Separate feature sets + exploiting target texts
  Label propagation less vulnerable to skewed class distribution than SVM wrapper algorithm

- **How does it compare to MT based methods?**

  English, Swedish: LP > MT
  Chinese: MT > LP
  Future work: Integration of MT based features into LP algorithm,
  Extension to different language pairs and genres

Thank you!