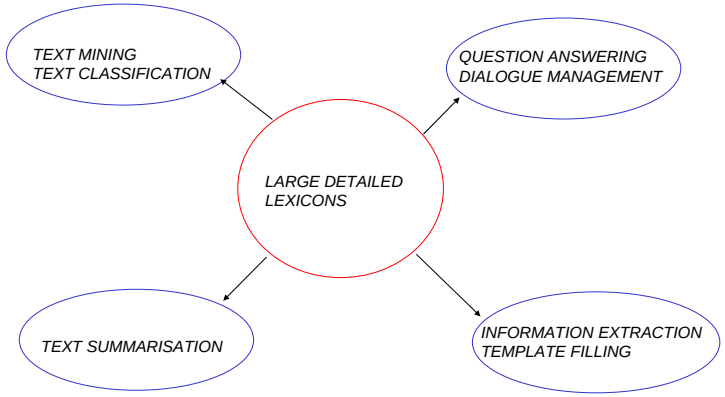
MULTI-LINGUAL PARADIGMS FOR AUTOMATIC LEXICAL ACQUISITION

Paola Merlo
University of Geneva

Why automatic lexical acquisition?



Why Multi-lingual?

- Accurate automatic lexical learning requires both
  - a well-founded theory of lexical representation
  - a distributional analysis of language

- Multi-linguality provides
  - abstract, general level of linguistic description
  - more data

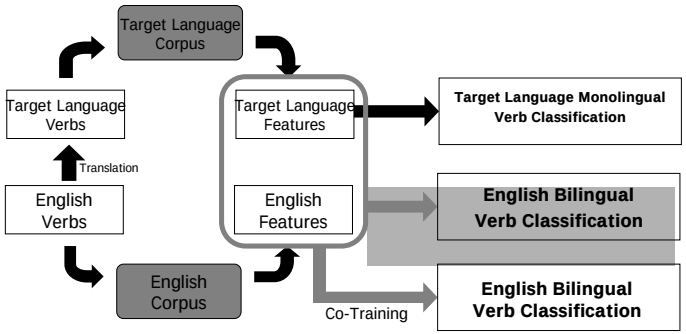
➡ Greater **coverage** and **accuracy** are possible by looking at several languages

Multi-lingual Generalisations

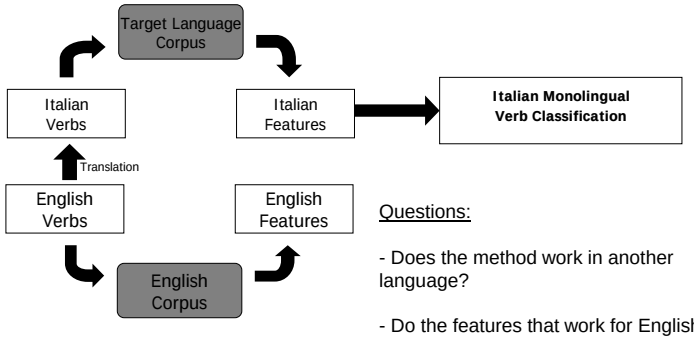
- Extension of mono-lingual method to a new language (Italian)
  - Shows **similarities** in the relations between frequency distributions and thematic relations across languages
  - Extends **coverage** to new languages

- Extension to the use of multi-lingual data to classify verbs in a given language (Chinese and English data to classify English verbs)
  - Shows that surface **differences** across languages are related to similar underlying meaning components
  - Improves **accuracy** in the classification of a given language

Three Experiments



Monolingual Italian Verb Classification

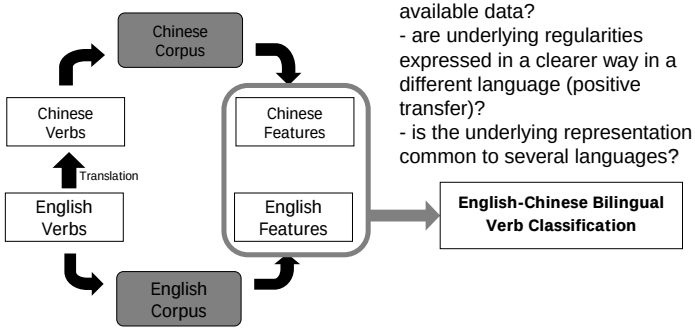


Questions:

- Does the method work in another language?

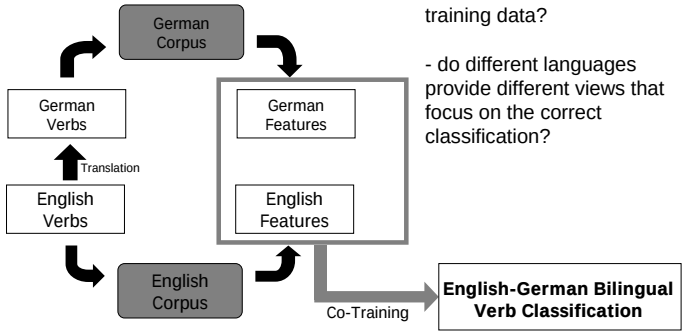
- Do the features that work for English also work in another language?

Bilingual Supervised English-Chinese Classification



Questions
- can we increase the amount of available data?
- are underlying regularities expressed in a clearer way in a different language (positive transfer)?
- is the underlying representation common to several languages?

Bilingual Semi-supervised English-German Classification



Questions
- can we reduce the amount of training data?

- do different languages provide different views that focus on the correct classification?

Multi-lingual Generalisations

Extension of mono-lingual method to Italian

Le classi di verbi

Modo di movimento	**Il fantino fa correre il cavallo.** ***Il fantino corre il cavallo.** **Il cavallo corre.**
Cambiamento di stato	**Il cuoco scioglie il cioccolato.** **Il cuoco fa sciogliere il cioccolato.** **Il cioccolato si scioglie.**
Performance	**Il bambino canta una canzone.** **Il bambino canta.** ***La canzone canta.**
Psicologici	**Gianni ama i soldi** **Gianni ama.** ***I soldi amano.**

(Alcune) Proprietà delle classi (e loro frequenza d'uso)

	Uso Transitivo	Uso Causativo lessicale	Animatezza Soggetto	Aspetto
MoM	bassa/nulla	Bassa/nulla fare	alta	attività
CoS	media	alta	bassa	achievements
Perf	alta	bassa	alta	attività
Psy	alta	bassa	alta	stati

Indicatori per la classificazione automatica

Gli indicatori per la classificazione automatica sono correlati di queste proprietà linguistiche, tradotte in fenomeni di superficie che possono essere contati facilmente in un corpus.

Indicatori per la transitività

$TRANS_v$:
- **1 se il verbo è usato transitivamente**
- **0 se il verbo è usato intransitivamente**

$PASS_v$:
- **1 se il verbo è di forma passiva**
- **0 se il verbo è di forma attiva**

$PartP_v$:
- **1 se il verbo è participio passato**
- **0 se il verbo non è participio passato**

Altri indicatori

Causatività	% intersezione soggetto-oggetto
	% uso di fare
Animatezza	% soggetto animato
	% oggetto animato
Aspetto	% gerundivi
	% avverbi
	% terminare/smettere
Ausiliare	% avere
	% essere

Il corpus

- Corriere della Sera
- Il Sole 24 Ore
- 84 milioni di parole
- etichettato con le parti del discorso
  (ringraziamo Achim Stein, Stoccarda)
  ma non analizzato sintatticamente

I conteggi delle frequenze sono tutti approssimati

I calcoli nel corpus: transitività

% transitivo

% passivo

% participio passato

Problema: individuare l'oggetto

Particolarità: uso dei clitici

Validazione: transitività

VERBI	CLASSI	DATI AUTOMATICI	DATI A MANO
Adorare	PSY	0.83	0.92
Apprezzare		0.81	0.92
Avanzare	COS	0.63	0.55
Cambiare		0.32	0.58
Arrangiare	PERF	0.89	0.16
Comporre		0.37	0.68
Balzare	MOM	0.17	0.00
Affrettare		0.92	0.10

I calcoli nel corpus: causatività

% intersezione soggetto-oggetto

L'oggetto della causativa transitiva è lo stesso argomento del verbo che il soggetto della forma intransitiva. Approssimiamo questa nozione come segue

- Estraiamo l'insieme di parole oggetto e l'insieme di parole soggetto
- Calcoliamo l'intersezione dei due insiemi
- Calcoliamo la proporzione dell'intersezione sulla somma dei soggetti e degli oggetti

Problema: soggetto nullo

% uso di fare

I calcoli nel corpus: animatezza

Approssimazione coi pronomi personali

% soggetto animato

problema: soggetto nullo

% oggetto animato

Particolarità: Uso dei clitici

I calcoli nel corpus: aspetto

Aspetto

% gerundivi

individua i verbi di stato

% avverbi

procedimento elaborato per l'inglese

% terminare/smettere

Gli esperimenti

Materiali 40 verbi per classe, 4 classi

Performance di base 25% di verbi classificati correttamente

Metodo Algoritmo di induzione: C5.0 (alberi di decisione)
Addestramento/Test: 30 verbi addestramento, 10 test

Vettore: [verbo,TRANS,PASS,PartP,CAUS,ANIM,…., classe]
Esempio: [aprire, .69, .09, .21, .16, .36, …., CoS]

Risultati

- Risultati globali: **57.5%** usando solo due famiglie di indicatori
  TRANS e ASPetto

- Commenti:
  l'indicatore più discriminante (radice dell'albero):TRANS
  al secondo livello e al terzo livello
  troviamo PASS, PartP, GERundio, Avverbi

  Tutti gli indicatori sono utili

  La classe meglio classificata: CoS (8/10)
  Le altre pari merito: 5/10

Discussione

- Performance soddisfacente (riduzione a metà del tasso di errore)

- Indicatori molto « rumorosi » (esperimenti precedenti con tratti classificati a mano arrivano al 86.5% accuratezza)

- Soggetto nullo crea grossi problemi all'indicatore animatezza

- Nuovi tratti per l'italiano non fanno molto

Conclusione

- Il metodo generale elaborato per l'inglese si applica all'italiano

- Problema di precisione dei tratti

- Apprendimento con pochi tratti grazie all'analisi linguistica

- I tratti più utili sono gli stessi che per l'inglese

Scoperta Non c'è bisogno di creare nuovi tratti per ogni classe

Interesse pratico Iniziare una classificazione per una lingua che non ha classifficazione verbale stabilita e nemmeno corpora analizzati

English-Chinese Bilingual Supervised Verb Classification

Problem

Work on Italian feature animacy shows that sometimes it is difficult to develop clear, easy-to-count surface indicators in a given language.

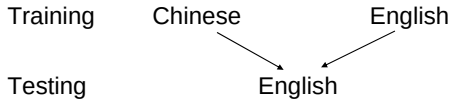
Taking Advantages of Cross-language Differences

(Tsang, Stevenson and Merlo, COLING 2002)

What is implicit in one language might be explicit in another

e.g. - Psych verbs in German often have a *sich* reflexive form
- Causative forms in Chinese are morphologically marked

Data from several languages classify one language



Mono-lingual Classification using Multi-lingual Data

English verb classes: MoM, CoS, C/T, 20 verbs in each class
English features: TRANS,PASS,VBN,CAUS,ANIM.

Chinese translations of the verbs (several translations are kept)

Counts of new features adapted to Chinese

- POS tags (indication of subcategorization and stative/active)
- passive particle
- periphrastic causative particle

Overt expressions of semantic information that is not overtly expressed in English

Materials and Method

English data from BNC (100 million words, tagged and chunked)
Chinese data from Mandarin News (165 million characters)

Vector template: [verb, Eft1,Eft2,…,Eftn, Cft1,Cft2,…,Cftn, class]

Learning: C5.0 (decision tree induction)
Training/Testing: 10-fold cross-validation 50 repeats

Results

Features	Acc%	SE%
Best English : ANIM,TRANS	**67.6**	**0.4**
Best Chinese : CKIP	**82.0**	**0.3**
Best combination : ANIM,TRANS,CKIP	**83.5**	**0.5**

- Best result: combination of Chinese and English features
- Chinese POS tags outperform the available English features

Discussion

- Surface differences across languages provide different views of the same data to the learning algorithm
- Underlying commonalities are confirmed, as different views converge on same classification
- Practically, considerable increase in amount of available data

Similarity to L1:L2 transfer

•Inspired by the phenomenon of transfer in L2 acquisition: beginner learners show facilitation when the patterns of their native L1 are coherent with L2 (positive transfer), while they show interference (negative transfer) when their L1 differs

Example
•German L1 auxiliary manner motion *sein* → L2: English = NT
•Chinese L1 difference between whisper and talk → L2 English= PT

Question: What is the *mechanism* that supports transfer?

Previous work could indicate that transfer occurs by indicators based on frequencies of usage. No need to have parallel corpus.

English-German Bilingual Semi-supervised Verb Classification

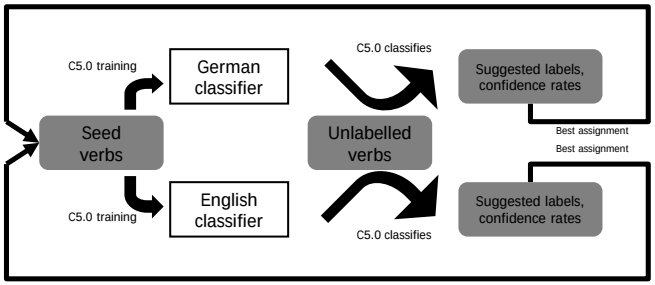
Problem

Previous work on bilingual classification requires large amounts of labelled data

Labelling data is costly and error-prone

Can we reduce the amount of labelled data using multi-lingual resources?

Co-Training



English- German Bilingual Classification Using Co-Training

Few annotated examples but a lot of unlabelled data

Two independent and sufficient views of the data

Well-suited to the situation of bilingual classification

Same importance attributed to German and English

The Data - Verbs

The Verbs

160 English verbs — Four verb classes
20 training and 20 test verbs per class

Verb Classes (optionally transitive)

Psychological Verbs:	*amuse, provoke, reassure*
Change of State Verbs:	*accelerate, broaden, melt*
Manner of Motion Verbs:	*climb, swim, walk*
Build and Performance Verbs:	*arrange, compose, cook*

Comment on Translation

No idioms or multi-word expressions
Manual verb class disambiguation
Validity only with back-translation or *Wordnet* - match

The Data - Corpora

German: The European Language News Corpus (1995)
Deutsche Presseagentur
90 million words
Unannotated
Chunker used: *Tree Tagger* and chunker (STTS tagset) by Helmut Schmid (1994) and (1995)

English: The British National Corpus (1995)
News articles, fiction, scientific texts
100 million words
Chunked (CLAWS POS tagset)

The Data - Features

English

Syntactic Feature Groups:

Transitivity	(2)
Animacy	(9)
Causativity	(2)
Prepositions	(9)
Preposition Groups	(3)
Grammatical Functions	(3)
Overlap	(12)
Others	(6)

German

Syntactic Feature Groups

Transitivity	(3)
Animacy	(2)
Causativity	(1)
Prepositions	(9)
Preposition Groups	(4)
Others	(2)

Morphological Feature Groups

Prefixes	(4)
Aspect	(3)

Results of Bilingual Co-Training

	Co-Training	Reference Trees:		
		German	English	Both
All features	59.4%	44.6%	62.4%	60.2%
Basic Features	59.0%	42.4%	64.5%	62.7%
Best group combinations on bilingual data German: Transitivity, Prepositions, Prep.Gr., Prefix English: Causativity, Animacy, Prepositions, Others	59.0%	45.4%	56.7%	54.3%

2-3 verbs

Method: C5.0 (decision tree induction)
10 seed verbs, 4-fold cross-validation, repeated 10 times

Discussion

No clear improvement in accuracy using co-training

Best feature set is more stable using co-training

Conclusions

- Automatic verb classification is a viable method to address the lexical acquisition bottleneck as it can be performed based on simple corpus counts

- Results from using cross-linguistic similarities and differences show
  - practical extensibility of method to new languages
  - satisfacotry performance but room for improvement
  - increase in coverage and performance using multi-lingual corpora
  - no clear improvement using semi-supervised techniques

- Results demonstrate combination of deeper linguistic knowledge with the robustness and scalability of statistical techniques.

Thank you