Computational Experiments on Verb Classes

what works and what doesn't

Paola Merlo
University of Geneva

Saarbrücken, 28.02.05

Collaborative work

Suzanne Stevenson Toronto

James Henderson Edinburgh

Eva Esteve Ferrer Geneva now Sussex

Eric Joanis   Toronto now Geneva

Michael Röösli Geneva

Vivian Tsang Toronto

Why do we care about classes of words in NLP?

- Automatic lexicon construction, extension, maintenance
  - words can be organised around shared syntactic and/or semantic properties
  - consistent extension

- Efficiency (smaller lexicon)
  – E.g. Experience of LexOrg/LexTract (Fei Xia)

- Class-based back-off or smoothing
  – Classes provide a level of more abstract even to collect counts

Why verb classification ?

Verbs are the primary source of relational information in a sentence

Jane hit the ball

NP NP

Agent Theme

For labelling tasks: argument structure, theta role labelling.

For structure building tasks: parsing, machine translation.

For information management tasks: information extraction, text mining

Types of classification

Syntactic information -- subcategorization frames

- Lapata 99,McCarthy and Korhonen 98

Semantic information

- selectional restrictions (Resnik 96)
- verbal aspect (Siegel and McKeown 01)
- lexical semantic classes (Aone and Mckee 96,
  Merlo and Stevenson 01, Joanis 02,
  Lapata and Brew 04, Schulte im Walde 03,
  Esteve Ferrer 04, Boleda 04)

Example of verb classification

English verb classes according to Levin

approximately 200 classes for 3000 verbs

For example

Manner of Motion:	race, jump, skip, moosey
Sound Emission:	buzz, ring, crack
Change of State:	burn, melt, pour
Creation/Transformation:	build, carve
Psychological state:	admire, love, hate, despise

Inductive application of Levin's hypothesis

Verbs which share semantic properties also share syntactic properties

There is a regular mapping from meaning components to syntactic usage (Levin 93, Pinker 89)

Can reason in reverse direction and induce semantic class from syntactic usage

⬇

Learn verb classes based on semantic properties using only corpus-based statistics

Two problems related to verb classes

Token-wise verb classification (verb class disambiguation)

for one ambiguous verb occurrence,

assign it to class

Mark *smashed* a fist on the desk in a defiant gesture (HIT)

I just wanted to go out and *smash* a door down (BREAK)

Type-wise verb classification

for a given lexical entry assign it to a class

RUN MoM or Cos or Psych?

Token-wise verb classification: two steps

For an ambiguous verb occurrence,

determine the probability to belong to a class in general (prior probability)

➡ based on subcategorisation frames

modify the general probability to take the current context into account (posterior probability)

➡ based on lexicalised alternations

Prior probability

Subcategorisation frames are very informative on the verb class (Lapata and Brew 04)

GIVE NP V NP NPto, NP V NP NP

PERFORMANCE NP V, NP V NP, NP V NP NP, NP V NP NPto, NP V NP NPfor

Lapata and Brew 04

$$P(v, f, c) = \cancel{P(v)}P(f \mid v)P(c \mid f, \cancel{v})$$

$$\approx P(f \mid v)P(f \mid c)P(c)$$

Our (HJM)

$$P(v, f, c) = \cancel{P(v)}P(c \mid v)P(f \mid c, \cancel{v})$$

$$\approx P(c \mid v)P(f \mid c)$$

Classes and Alternations

Spray/Load verbs

I loaded ***hay*** *into the wagon*.

I loaded the wagon ***with hay*** .

Run verbs

The jockey jumped the horse ***over the fence***

The horse jumped ***over the fence***

The horse jumped ***the fence***

Model of alternations

Modelling alternations directly requires calculating the probability of the sentences with which the current sentence could alternate in the text.

This is a model of context where context is not defined by string adjacency, but it is a linguistic paradigm.

0 *The jockey*$_j$ *jumped the horse*$_k$ ***over the fence***$_l$

i *The horse*$_j$ *jumped* ***over the fence***$_k$

i+1 *The horse*$_j$ *jumped* ***the fence***$_k$

$$a_{ij0k} = \langle POS, Sij, S0k \rangle$$

$$a_{iji+1k} = \langle NEG, Sij, Si+1k \rangle$$

$$a_{i+1k0l} = \langle POS, S_{i+1k}, S_{0l} \rangle$$

A NEG label also applies to alternating slots of different senses of the same verb

Model of alternations

We assume independence of sentences,

independence of slots

and independence from the verb given the class

(i.e. all verbs in a class behave homogeneously).

$$\prod_{a_{ij0k} \in A} P(t \mid \langle Sij, S0k \rangle, f_i, f, c, c_i)$$

Model of alternations

$$\prod_{a_{ij0k} \in A} P(t \mid \langle S_{ij}, S_{0k} \rangle, f_i, f, c, c_i)$$

Two cases: $c = c_i$, it is a true alternation

c not $= c_i$ noise in an ambigous verb, the alternating slots belong to different senses and are overlapping by chance

We estimate both cases calculating the overlap of slots of unambiguous verbs. For the model of noise we generated data artificially. If unambiguous verbs not sufficient, we assume a uniform distribution over classes of ambiguous verbs.

Experimental Materials

Corpus: British National Corpus (parsed with Henderson 2003)

Two data sets

Our 40 verb occs each from 5 classes (psych, cos,mom,ben,spray/load)

100 random verb occs

117 frequency stratified occs

After filtering, 370 occurrences, hand annotated for correct class

LB04 1840 occurrences hand annotated

(datives, benefactives, possessives and conatives)

Results

	Data Sets	
	OUR	LB04 (part of)
N	370	1840
Random baseline	47.8%	42.1%
LB04 prior	48.6%	**46.4%**
Prior w/o alts	**54.6%**	45.3%
Posterior with alts	50.2%	43.4%

Similar results even if we use back-off to Wordnet classes for the heads of slots on which we calculate the overlaps

Results by class

OUR DATA SET					
	psych	cos	run	ben	spray/load
N	37	36	32	36	33
Random baseline	35%	36%	34%	37%	41%
LB04 prior	16%	61%	38%	47%	24%
Prior w/o alts	51%	53%	**59%**	<u>42%</u>	58%
Posterior with alts	43%	36%	**62%**	<u>56%</u>	33%

Conclusions

Alternations do not provide useful information beyond set of subcategorisation frames

Alternations are difficult to model properly and to estimate

Properly modelled prior much more useful.

Conjecture: alternations provide a notion of context that is too wide. This conjecture is supported by negative results in LB04, who also found that collocations do not help but narrowly defined context (small windows of words) does help in disambiguation.

Verb Classification

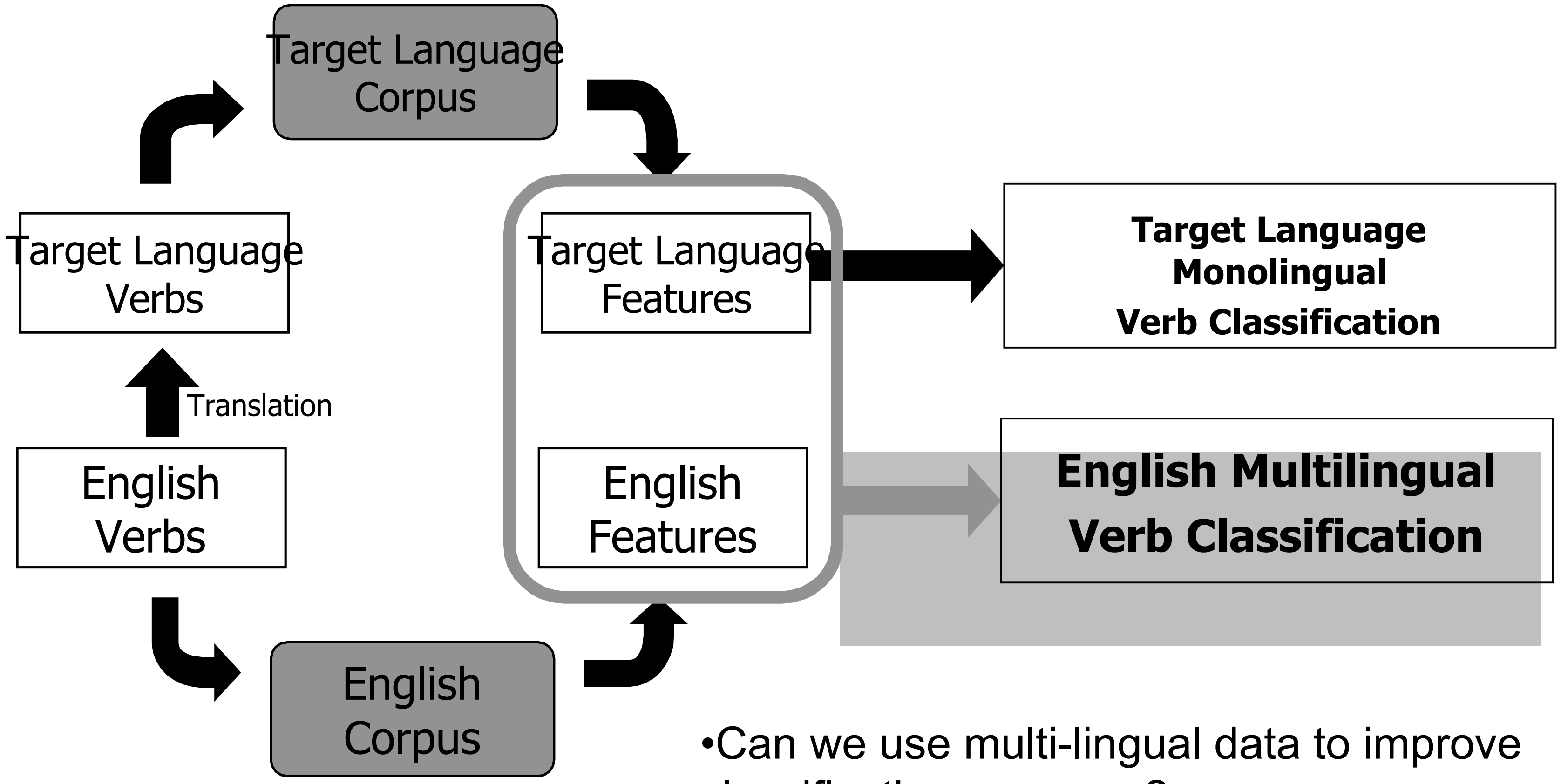
In the solution of the token-wise verb disambiguation, we need to know among which classes we need to choose.

But even for unambiguous verbs, sometimes the syntax is not unambiguously telling us how to assign the verb to the class.

Problem: n-class forced choice

Given a set of syntactically very similar classes, determine to which class verb belongs.

Experiments



•Can we use multi-lingual data to improve classification accuracy?

English Optionally Intransitive Verb Classes

Manner of Motion	**The rider** *(Causal) Agent*	**raced**	**the horse** *Agent*	**past the barn**
	The horse *Agent*	**raced**	**past the barn**	
Change of State	**The cook** *(Causal) Agent*	**melted**	**the butter** *Theme*	
	The butter *Theme*	**melted**		
Creation/Transformation	**The contractors** *Agent*	**built**	**the house** *Theme*	
	The contractors *Agent*	**built**	**all summer**	

Methodology **(Merlo Stevenson 2001)**

Analyse verb classes to determine discriminating thematic properties

Develop indicators that approximate thematic properties and that can be counted in a corpus

Collect relative frequencies to generate a statistical summary of the thematic behaviour of each verb

Apply machine learning algorithm (e.g. decision tree induction) to produce a classifier

The Basic Idea

Underlying abstract differences among the verb classes will surface as detectable differences in the usage of surface indicators

Classes	Transitive		Intransitive
	Subject	Object	Subject
Manner of Motion	(Causal) Agent	Agent	Agent
Change of State	(Causal) Agent	Theme	Theme
Performance	Agent	Theme	Agent

Transitive Use

- **Transitivity by causation is more complex**
- **Agent object is (typologically) rare**
- **MoM < CoS < C/T**

Animacy of Subject

- **Themes are less likely to be animate**
- **CoS < {C/T,MoM}**

Initial English Supervised Experiments

Materials - 59 verbs (20 MoM, 19 CoS, 20 C/T)
- 65 million tagged words (29 million parsed) (WSJ and Brown corpus)
- BNC 100 million tagged words

Features Estimated by simple relative frequencies

Vector template: [verb,TRANS,PASS,VBN,CAUS,ANIM,class]
Example: [open, .69, .09, .21, .16, .36, CoS]

Method Learner: C5.0 (decision tree induction algorithm)
Training/Testing: 10-fold cross-validation repeated 50 times

Results

Overall results: accuracy 69.8% - 82.4% (baseline 33.9, expert upper bound 86.5%)

large reduction in error rate on previously unseen verbs

Effectiveness of features

All features, except PASS, are useful in classification

Analysis of errors

Hypothesized relation between features and thematic assignments is confirmed

Extension to new classes

(Joanis and Stevenson 2003)

Generic feature space extending linguistically defined space

- Syntactic slots (120 features)
- Tense, voice, aspect (24 features)
- Animacy (76 features)

Very good results

- 23-47% over baseline
- 40-70% reduction in error rate

In most cases, generic feature space does as well as when linguistic expertise is involved in selecting features.

Conclusion

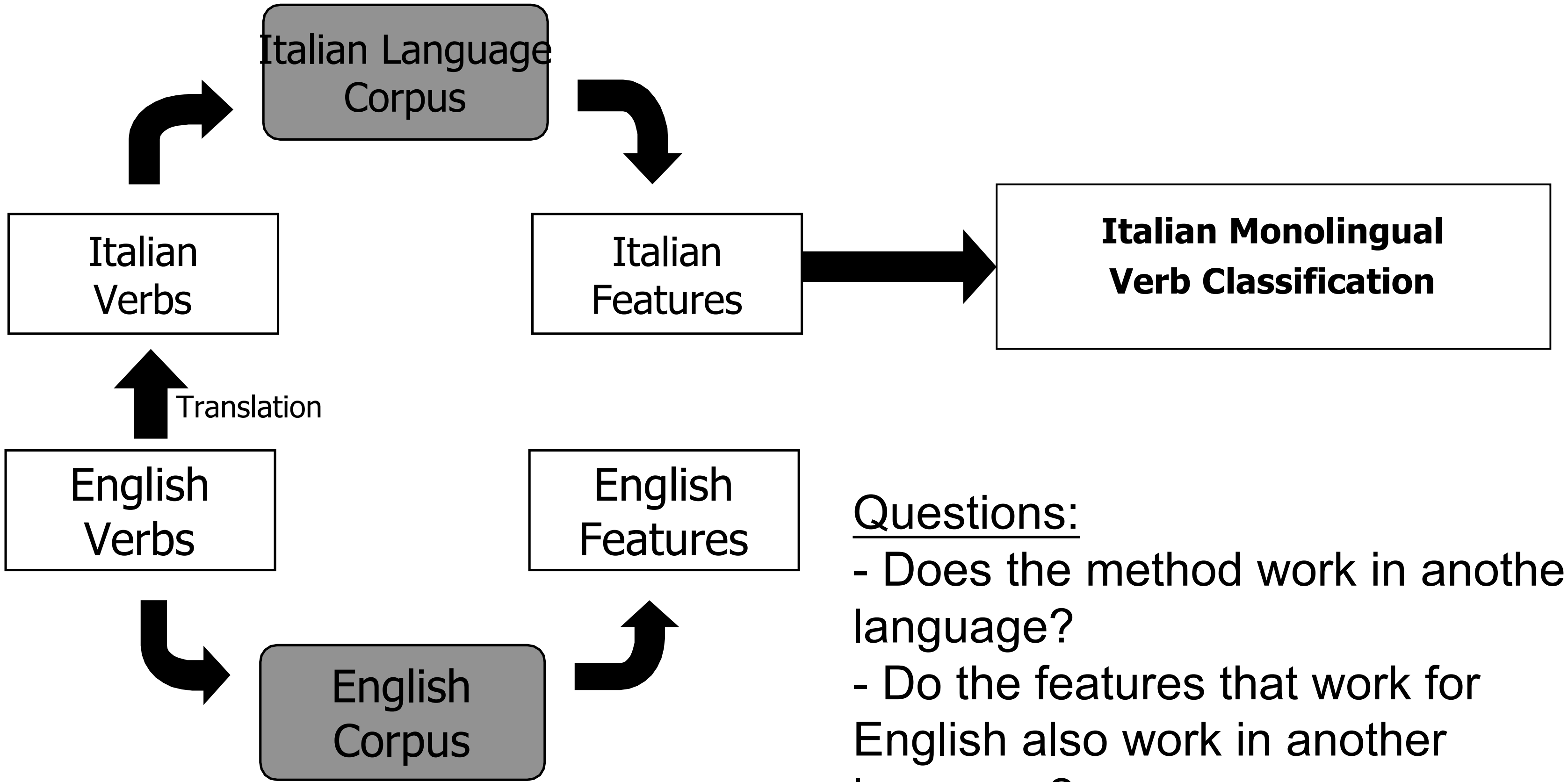
Hypothesis confirmed

corpus-based indicators reflect underlying
semantic properties of verbs

Method can have high performance

Discovery We do not need to investigate new indicators
for each new class.
Indicators can extend to a more general form for
for many classes with very good results
(Joanis and Stevenson 2003,
Merlo and Esteve Ferrer 2004).

Monolingual Italian/German Verb Classification



Questions:

- Does the method work in another language?
- Do the features that work for English also work in another language?
- How do we develop a new feature space?

Extension to Italian: Feature Space

Features inspired from English: Transitivity,
Animacy,
Causativity

New Aspectual Features

Potential problems: null subject (very frequent)
flexible word order (postverbal subject)

Potential new language specific features: clitics
auxiliary selection

Results

Results: **57.5%** (baseline 25%)

using only two groups of indicators

Transitivity and aspect

Comments:

most discriminating feature (root of the tree): TRANS

second and third level

PASS, PartP, GERundive, Adverbs

All features are useful

Class classified with best accuracy: CoS (8/10)

Discussion

Reasonable performance (40% reduction of error rate)

Very noisy features (previous pilot experiments with hand collected animacy had reached 86.4% accuracy)

Null subject creates big problems

New language-specific features (auxiliary selection) do not do much

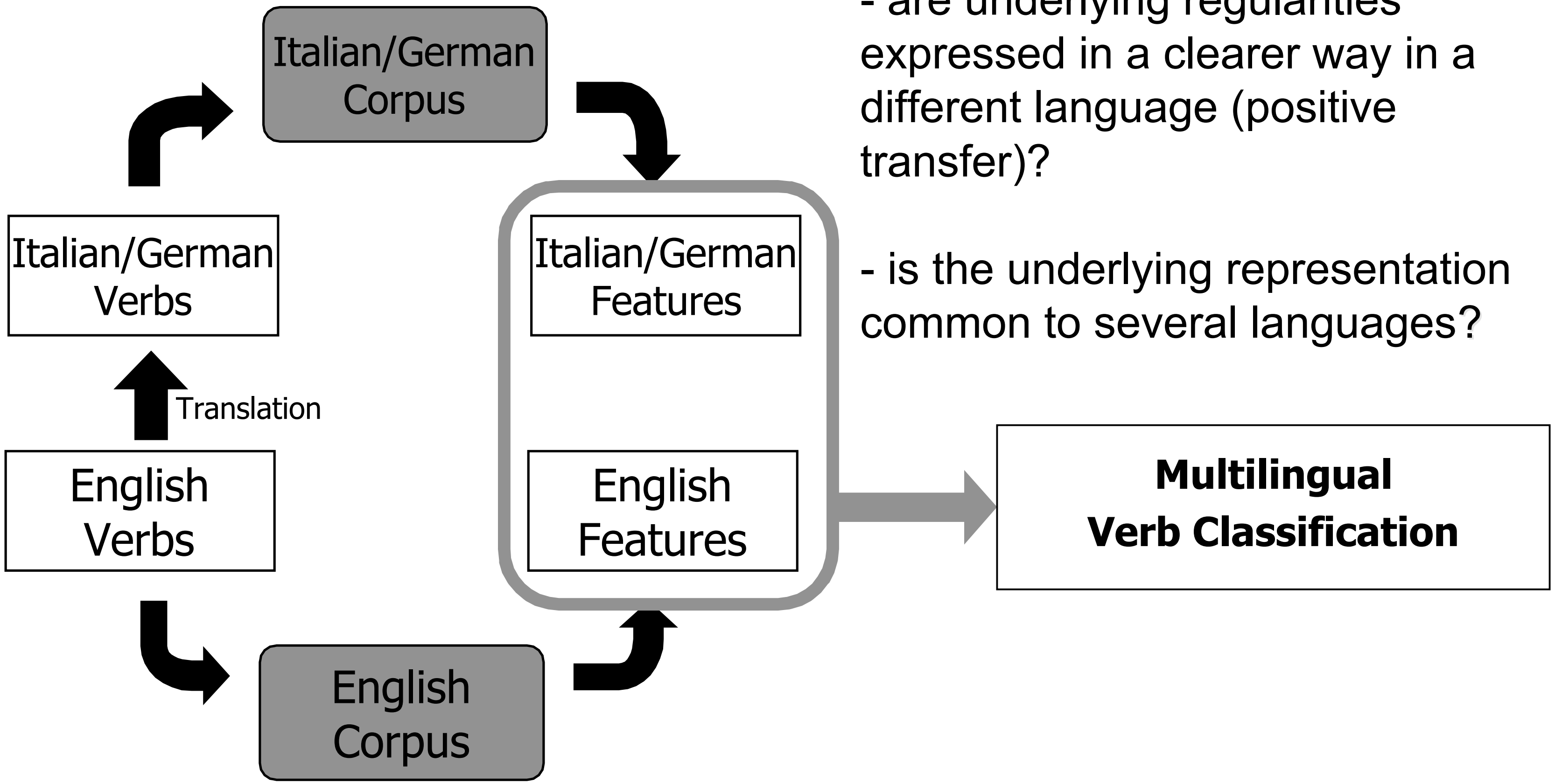
Practical Interest Bootstrap a classification in a new language for which there is no existing classification

German Results (Röösli 2004)

Baseline:	25.0%
Basic Features:	48.8%
Best combination:	53.8%

Error reduction of up to 38.4%

Multilingual 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?

Taking Advantages of Cross-language Differences (Tsang 2001,Tsang, Stevenson and Merlo 2002, Stevenson, Merlo, Tsang forthcoming)

What is implicit in one language might be explicit in another

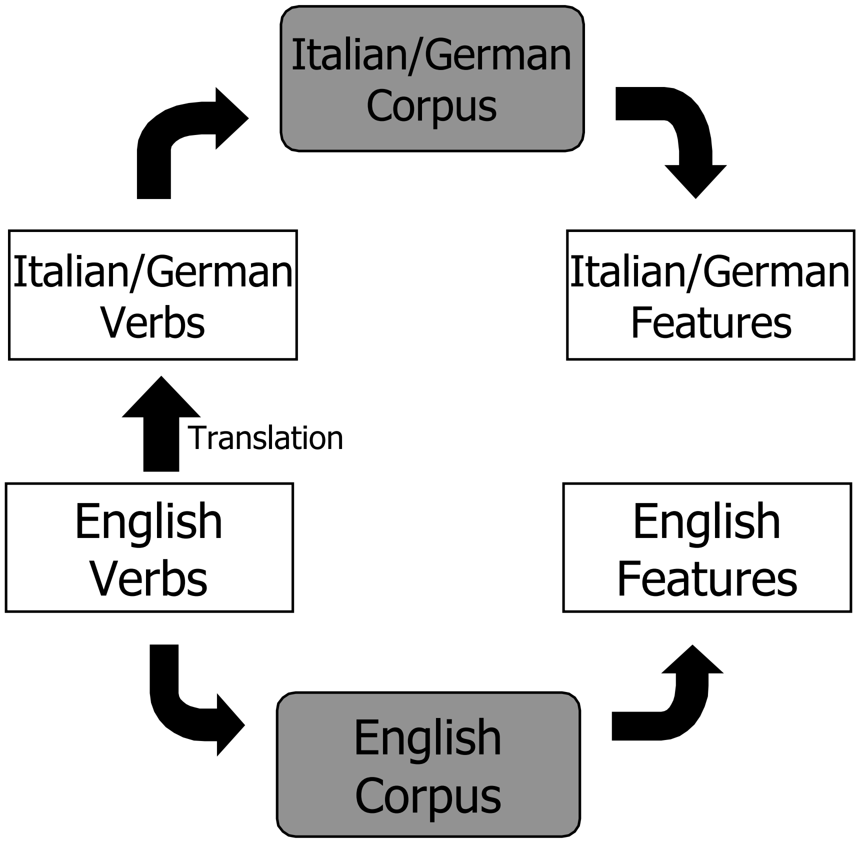
e.g. - Psych verbs in German often have a *sich* reflexive form
- Causative forms in Chinese are morphologically marked

Data from several languages classify one language

Training Chinese English

Testing English

Multilingual aligned vectors



Translate English verb

For all translations, back translate and keep those verbs whose back-translation contains initial English verb

No need for parallel corpus

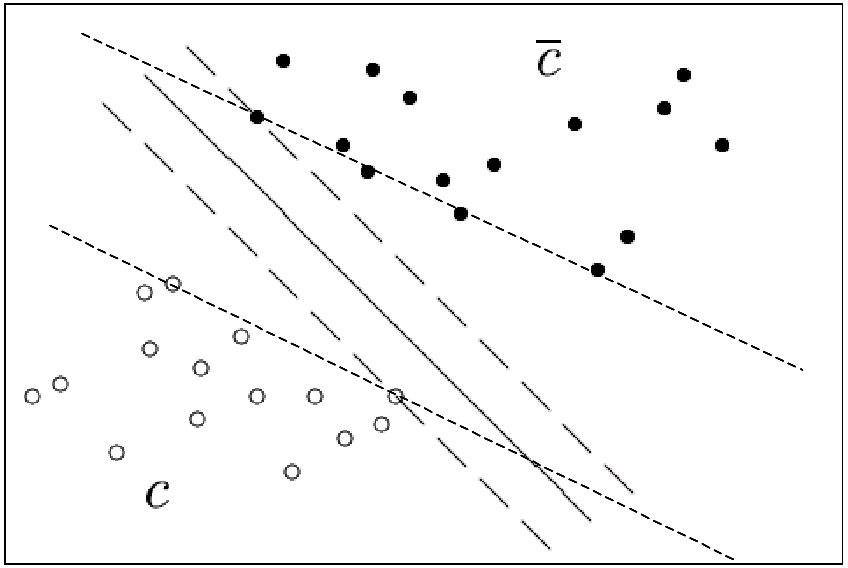
Target language features are an average of all translations

Vector: [verb, <English FTs>, <Italian Fts>, <German Fts>, class]

Support Vector Machines

SVMs assume that only border cases (the support vectors that define the margins between two classes) really matter in classification, and try to find the largest margin between two classes.

Sometimes this requires transforming the space into a higher dimensional space to be able to separate the classes with a linear function.



Multi-lingual preliminary results (work in progress)

	E=general feature space	E= Levin derived
E	81%	76%
EI	80%	82%
EG	81%	80%
EIG	76%	90%

- General feature space better than hand picked
- Too many features and too many languages confuse learner
- Task specific features provide views and generate better performance
- Top task specific better than generic

General comments

Verb categorisation can be successfully performed based on unparsed text using only surface cues.

Main features are transitivity, animacy (related to thematic properties) and aspect. Features related to alternations are least useful.

Features transfer across languages.

Languages can provide different views on underlying common classification, improving accuracy.

Thank you