Development and evaluation of phonological models for cognate identification

Bogdan BABYCH

Centre for Translation Studies, University of Leeds, UK

b.babych@leeds.ac.uk

Approach

- Cognate identification important for Machine Translation (MT) & CAT:
  - Building cognate term-banks from comparable corpora;
  - Lexicon for closely-related & under-resourced languages;
  - Additional data source for statistical sentence alignment, etc...
- Current methods rely on character-based distances (e.g., Levenshtein)
  - Count insertions, deletions, substitutions of characters;
  - Treat each character as an 'atomic' unit without structure;
  - Difficult to apply across different scripts (need transliteration).
- Solution: phonological features of characters to calculate similarity
  - Characters mapped to sets of distinctive acoustic features;
  - Used in historical linguistics and dialectological studies;
  - Applied to cognate identification for MT (cf. Babych, 2016).

Graphemic-Phonological features
Uk: "жовтий" (zhovtyj) = 'yellow'

ж (zh) 'type:consonant', 'voice:ff-voiced', 'maner:ff-fricative', 'active:ff-fronttongue', 'passive:ff-palatal'
о (o) **'type:vowel', 'backness:back', 'height:mid', 'roundedness:rounded', 'palate:nonpalatalizing'**
в (v) **'type:consonant'**, 'voice:fl-voiced', 'maner:fl-fricative', 'active:fl-labial', 'passive:fl-bilabial'
т (t) 'type:consonant', 'voice:pf-unvoiced', 'maner:pf-plosive', 'active:pf-fronttongue', 'passive:pf-alveolar'
и (y) **'type:vowel'**, 'backness:front', **'height:closemid','roundedness:unrounded', 'palate:nonpalatalizing'**
й (j) 'type:consonant', 'voice:xm-sonorant', 'maner:xm-approximant','active:xm-midtongue', 'passive:am-palatal'

Feature representations for corresponding characters in Ru: "жёлтый" (zheltyj) = 'yellow'.

ё (io) **'type:vowel', 'backness:back', 'height:mid', 'roundedness:rounded'**, 'palate:palatalizing'
л (l) **'type:consonant'**, 'voice:lf-sonorant', ' maner:lf-lateral', 'active:lf-fronttongue', 'passive:lf-alveolar'
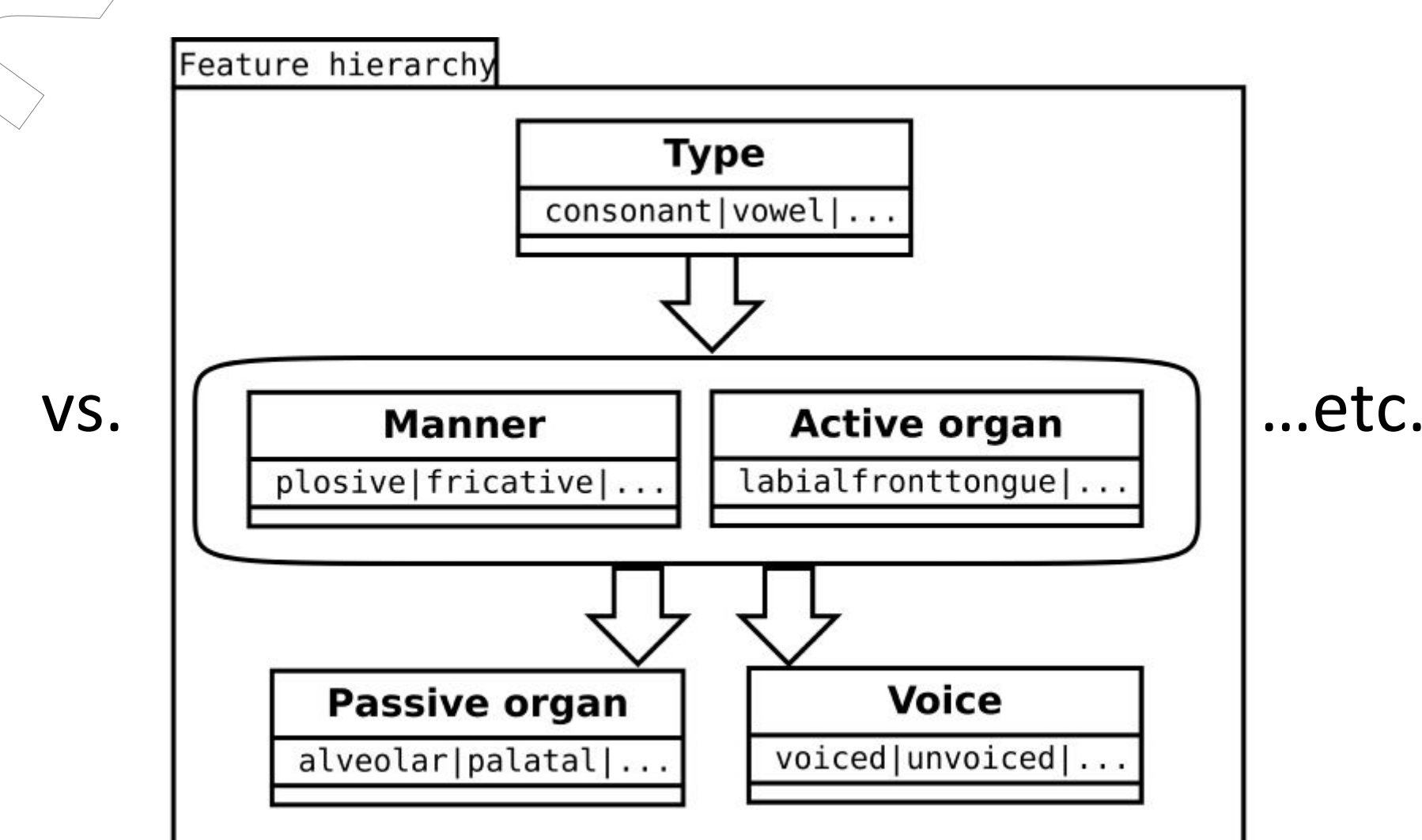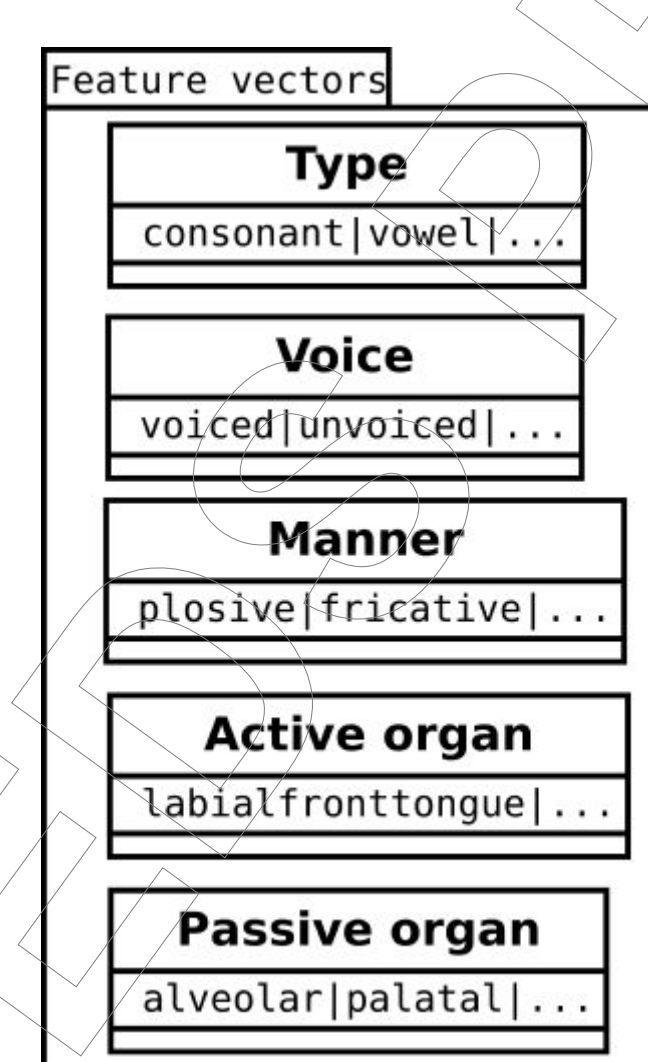ы (y) **'type:vowel'**, 'backness:central', **'height:closemid','roundedness:unrounded ', 'palate:nonpalatalizing'**

Calculation of the Phonological Levenshtein for Uk "жовтий" (zhovtyj) = 'yellow' and Ru "жёлтый" (zheltyi) = 'yellow':

```
0.0 1.0 2.0 3.0 4.0 5.0 6.0
1.0 0.0 1.0 2.0 3.0 4.0 5.0
2.0 1.0 0.2 1.2 2.2 3.2 4.2
3.0 2.0 1.2 1.0 2.0 3.0 4.0
4.0 3.0 2.2 2.0 1.0 2.0 3.0
5.0 4.0 3.2 3.0 2.0 1.2 2.2
6.0 5.0 4.2 4.0 3.0 2.2 1.2
```
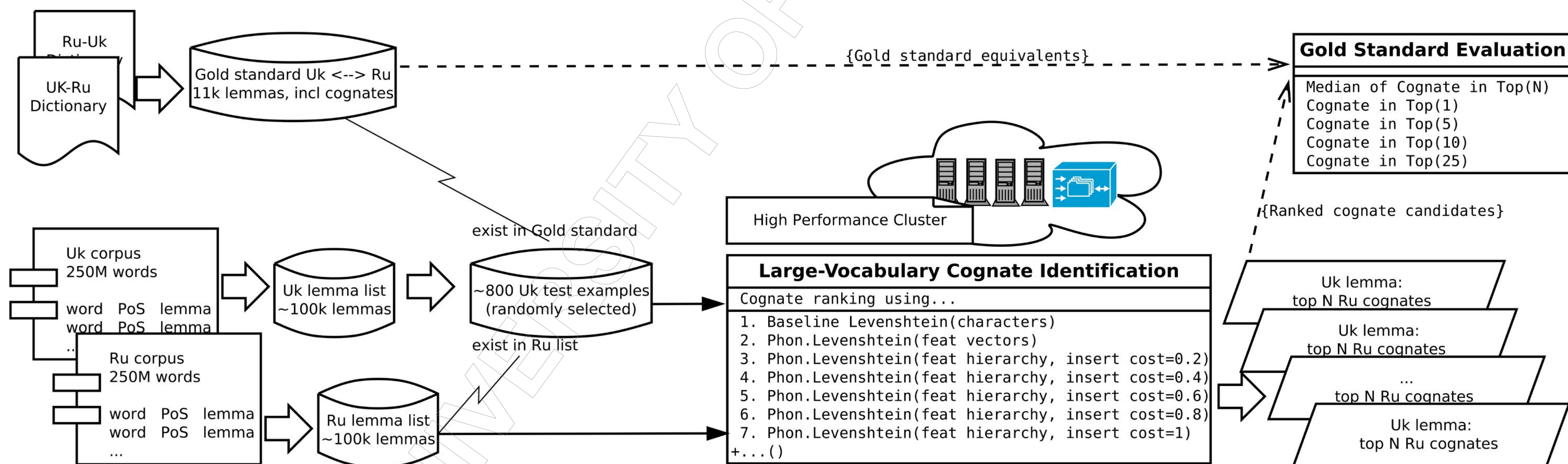
cf.: Metric calculated for
Uk "жовтий" (zhovtyj) = 'yellow' with
Ru "жуткий" (zhutkij) = 'dismal':

```
0.0 1.0 2.0 3.0 4.0 5.0 6.0
1.0 0.0 1.0 2.0 3.0 4.0 5.0
2.0 1.0 0.2 1.2 2.2 3.2 4.2
3.0 2.0 1.2 1.0 1.2 2.2 3.2
4.0 3.0 2.2 2.0 1.8 2.2 3.0
5.0 4.0 3.2 3.0 2.8 2.0 3.0
6.0 5.0 4.2 4.0 3.8 3.0 2.0
```

Problems and solutions

- Large-vocabulary cognate identification ≠ historic linguistic studies
  - False positives due to many more plausible alternatives;
  - In pilot: phonological feature vectors degraded performance;
  - Baseline character-based Levenshtein more resistant to errors.
- Some errors of Phonological Levenshtein metric due to imbalanced cost of insertion/deletion vs. substitution, e.g., for 5 phonological features:
  - Typical substitution cost for unrelated consonants = 0.8;
  - The insertion/deletion cost = 1.0 leads to under-generation of cognates with inserted/deleted characters.
- We need a systematic way for development and evaluation of alternative feature models, arrangements and edit costs (an automated framework for phonological feature engineering):

Feature vectors

| Type |
| consonant\|vowel\|... |

| Voice |
| voiced\|unvoiced\|... |

| Manner |
| plosive\|fricative\|... |

| Active organ |
| labialfronttongue\|... |

| Passive organ |
| alveolar\|palatal\|... |

vs.

Feature hierarchy

| Type |
| consonant\|vowel\|... |

| Manner | | Active organ |
| plosive\|fricative\|... | | labialfronttongue\|... |

| Passive organ | | Voice |
| alveolar\|palatal\|... | | voiced\|unvoiced\|... |

...etc.

Development and evaluation framework for phonological models

{Gold standard_equivalents}

Ru-Uk Dictionary
UK-Ru Dictionary

Gold standard Uk <--> Ru 11k lemmas, incl cognates

Uk corpus 250M words
word PoS lemma
word PoS lemma
...

Uk lemma list ~100k lemmas

exist in Gold standard

~800 Uk test examples (randomly selected)

exist in Ru list

Ru corpus 250M words
word PoS lemma
word PoS lemma
...

Ru lemma list ~100k lemmas

High Performance Cluster

Large-Vocabulary Cognate Identification

Cognate ranking using...
1. Baseline Levenshtein(characters)
2. Phon.Levenshtein(feat vectors)
3. Phon.Levenshtein(feat hierarchy, insert cost=0.2)
4. Phon.Levenshtein(feat hierarchy, insert cost=0.4)
5. Phon.Levenshtein(feat hierarchy, insert cost=0.6)
6. Phon.Levenshtein(feat hierarchy, insert cost=0.8)
7. Phon.Levenshtein(feat hierarchy, insert cost=1)
+...()

Gold Standard Evaluation

Median of Cognate in Top(N)
Cognate in Top(1)
Cognate in Top(5)
Cognate in Top(10)
Cognate in Top(25)

{Ranked cognate candidates}

Uk lemma: top N Ru cognates
Uk lemma: top N Ru cognates
...
top N Ru cognates
Uk lemma: top N Ru cognates

Evaluation results

Experiment	Median top N	Top 1	Top 5	Top 10	Top 25
Baseline Char.Lev	50	206	328	360	382
Phon.Lev FeatVectors	87.5	**215**	289	319	349
Diff with Base L	-75%	+4.4%	-10%	-11%	-9%
Phon.Lev Hierarchy:					
Phon.LevH i=0.2	125.5	216	291	315	342
Phon.LevH i=0.4	54.5	230	307	334	367
Phon.LevH i=0.6	*48*	235	328	354	385
Phon.LevH i=0.8	**40**	**240**	**337**	**359**	**391**
Phon.LevH i=1.0	47.5	**240**	334	359	385
Best Improv. over BaseL	**+20%**	**+16.5%**	**+3%**	**0%**	**+2%**

Conclusions

- Hierarchical phonological models improve accuracy of large-vocabulary cognate identification up to 20% on Top-N measures.
- Evaluation and development framework for phonological models:
  - Allows for accurate feature calibration and parameter setting;
  - Enables a systematic task-based feature engineering.
- Potential applications of the models beyond cognate identification:
  - Interlingual transliteration via common phonological space;
  - Character-based models for Neural MT;
  - Morphology induction & morphological variation modelling.
- Future work: evaluating alternative feature topologies, integrating statistical and semantic filters for cognate identification.
- Phonological models and resources (feature sets + scripts) released on https://github.com/bogdanbabych/cognates-phonology