CRC 1102 project C4: Modeling mutual intelligibility between Slavic languages

PIs: Tania Avgustinova, Dietrich Klakow, and Roland Marti
PhD stud.: Andrea Fischer, Klara Jagrova, and Irina Stenger

March 13, 2015

Motivation

We investigate **receptive multilingualism:** the passive understanding of a language without being able to use it productively.

As an example, consider a **Czech native speaker** reading a **Polish text** while **never having learned** Polish. How much information will this reader be able to decode? Which false information will they extract?

Our goal lies here: in formulating **measures of mutual intelligibility**.

We focus on **Slavic languages**: the richness and strong inter-relatedness of this language family makes it ideal for investigation. Additionally, a large body of traditionalistic linguistic knowledge is available to draw from.

C4's Mission Statement

1. Validate (or reject!) traditional assumptions about linguistic similarity and mutual intelligibility of Slavic languages;
2. Gain insights into differences of information en- and decoding between Slavic languages;
3. Correlate these differences with results from a series of experiments;
4. Provide differential models of information coding, and establish an empirical basis for applications which require these.

Contents of this Presentation

Edit Distance as Measure for Linguistic Distance

Most of the previous work on mutual intelligibility results in a very diffuse statement of *general* or *average* expected intercomprehensibility.

Typically, the average Levenshtein distance between pairs of cognates is computed and shown to correlate (well, actually) with human performance in intercomprehension tasks, such as direction giving or ad-hoc translation.

This approach has some drawbacks: it works only at the *language* level, disregarding any aspects that might make one text more or less comprehensible than another.

Can we somehow adapt this to work at a word-by-word level?

Language Models, Surprisal, and Reading Comprehension

Results from psycholinguistics consistently show that reading times correlate strongly with surprisal of *n*-gram language models.

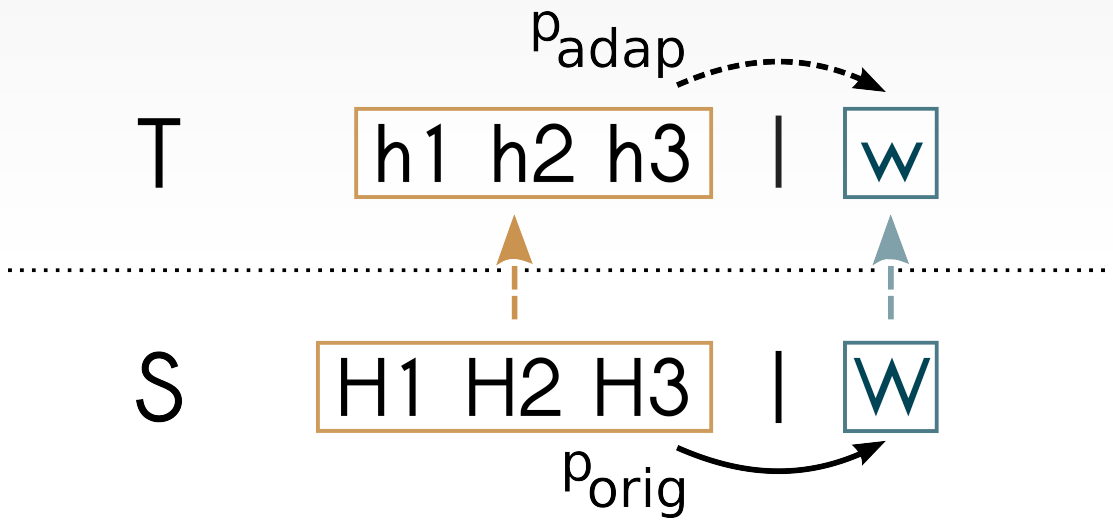
If we take reading time to be a gauge for *comprehensibility* of a text, then surprisal should also work as a measure of reading *inter*comprehension.

LM surprisal is the output variable we hope to correlate with human intercomprehension performance.

However, traditional language models cannot handle the differences between source and target language's en- and decoding schemes.

Language Model Domain Adaptation

This problem can be considered one of *domain adaptation*.



Formulaically:

$$p_{adap|W^s,H^s}(w^t|h^t) \sim p(w^t|W^s) \cdot p_{orig}(W^s|H^s) \cdot p(h^t|H^s)$$

Language Model Domain Adaptation - cont'd

The adaptation might be assumed to mean making hard decisions:

$$p(w^t|h^t) \approx \max_{W^s,H^s} p_{adap|W^s,H^s}(w^t|h^t)$$

or to mean making soft ones:

$$p(w^t|h^t) = \sum_{W^s,H^s} p_{adap|W^s,H^s}(w^t|h^t)$$

LM Domain Adaptation: Aspects

To the linguist, there are many different aspects of language to consider, such as orthography, morphology, lexis, syntax, and semantics.

To the LM, we only have to adapt

1. words of source to words of target language;
2. word order.

The second point could be considered solved: we can simply train statistical word alignments from parallel corpora.

Language Model Domain Adaptation - cont'd

One might utilize neural network or other LMs which operate with abstract word representations – e.g. word vectors or classes – for the task.

However, we are not really looking for abstract models which merely adapt, but rather for the *mechanisms* by which the languages code information.

We also hope to incorporate linguistic hypotheses to test – both of these requirements make models which are not readily interpretable infeasible.

A Multi-Class LM

Basing upon the classic "hard" class LM:

$$p(w|h) = p(w|c_w) \cdot p(c_w|c_h)$$

we relax the notion of singular classes to multiple ones:

$$p(w|w_h) = \sum_{f \in F(w)} p(w|f) \sum_{f_h \in F_\times(w_h)} p(f|f_h) \prod_{w_{h_i}} p(f_{h_i}|w_{h_i})$$

This model was recently investigated by Tobias Backes in the scope of his Master's thesis and is yet to be published.

A Multi-Class LM - cont'd

$$p(w|w_h) = \underbrace{\sum_{f \in F(w)} p(w|f)}_{\text{word is mixture of features}} \overbrace{\sum_{f_h \in F_\times(w_h)} p(f|f_h)}^{\text{feature paths can have different importance}} \underbrace{\prod_{w_{h_i}} p(f_{h_i}|w_{h_i})}_{\text{normalization}}$$

This multi-class, multi-path model allows us to:

- incorporate arbitrary features into word representations;
- analyze significant feature paths and their relative importance.

Feature Spaces and Projections

If we consider a word to be a *weighted* mixture of features, the domain adaptation can be done via *embedding* the features of one language into the space of features of the other.

Then, if the model normalizes with the original words, it doesn't necessarily do so with the embedded ones. As an example, consider Bulgarian and Russian: Bulgarian has no case declension. If we assume that every other feature has a perfect correspondent in Russian, Bulgarian looks like Russian where case is missing.

This would lead the adapted model to be, in effect, a *defective* probability distribution. The missing probability mass represents the uncertainty/noise in the intercomprehension process, and leads to a relative increase in surprisal.

Feature Spaces and Projections - cont'd

This leaves us with two tasks:

1. finding ideal in-domain features
2. finding ideal domain adaptation procedures

Both of these tasks lend themselves well to minimum description length: the nature of the model, being a probability density function, makes encoding data extremely straightforward.

The question then becomes which features and which adaptation procedures one needs to consider.

Possible Adaptation Procedures

Linguistically, typically one of two perspectives is taken:

1. direct correspondences
2. "common Slavic" model

Both of these might be inferred from Swadesh (cognate) lists.

Possible Adaptation Procedures - cont'd

In linguistic research, this is usually taken to mean transformation rules of the form $A \rightarrow B$, where A and B are simply substrings of the word under consideration.

Examples from Czech to Polish: $v \rightarrow w$, $\check{c} \rightarrow cz$, $\check{s} \rightarrow sz$

But also:
CZ provaz - PL powróz (rope)
CZ manželka - PL małżonka (wife)

It remains to be seen which types of rules are necessary to achieve coverage of cognate lists, and it is not clear yet to what extent there will be overlap between linguistic knowledge and automatically induced models.