Toward Pan-Slavic NLP:
Some Experiments with Language Adaptation

Serge Sharoff
Centre for Translation Studies
University of Leeds, Leeds, UK
s.lastname@leeds.ac.uk

1 Introduction

There is great variation in the amount of NLP resources available for Slavic languages. For example, the Universal Dependency treebank (**?**) has about 2 MW of training resources for Czech, more than 1 MW for Russian, while only 950 words for Ukrainian and nothing for Belorussian, Bosnian or Macedonian. Similarly, the Autodesk Machine Translation dataset only covers three Slavic languages (Czech, Polish and Russian). In this talk I will discuss a general approach, which can be called Language Adaptation, similarly to Domain Adaptation. In this approach, a model for a particular language processing task is built by lexical transfer of cognate words and by learning a new feature representation for a lesser-resourced (recipient) language starting from a better-resourced (donor) language. More specifically, I will demonstrate how language adaptation works in such training scenarios as Translation Quality Estimation, Part-of-Speech tagging and Named Entity Recognition.

2 Transfer of Feature Representation

Machine Learning algorithms are limited by the availability of training data. This problem is often addressed by developing algorithms to transfer NLP models across different domains, for example, an opinion mining model trained on IMDb can be transferred to the domain of hotel reviews (**?**). In a similar way, we can assume that a model trained in a donor language can be transferred to a recipient language relying on the fact that both languages come from the same language family.

One of the observations for transferring models across languages is that while the general assumption of similarity holds, the individual features exhibit a slightly different distribution. For example, in the task of estimating MT quality without reference translations, good MT examples are similar in the feature space describing translation into two related languages, but the exact feature values, such as the Language Model values or the phrase table sizes differ. One way of transferring the feature spaces is via Self-Taught Learning (STL), in which an autoencoder learns to reduce the dimensions of **unlabelled** datasets for the two domains. Then the available **training** set in one domain is transformed using the autoencoder, so that a new prediction model can be equally successful in the source domain and in the new target domain (**?**). As shown in (**?**), an application of this transformation to predicting the amount of Post-Editing needed to improve raw MT output can produce models which almost reach the accuracy of the original prediction model (Table **??**).

Upper baseline (ru)	MAE	0.18
	RSME	0.27
	Pearson	**0.47**

en-ru	→	en-cs	en-pl
STL	MAE	0.19	0.19
	RMSE	0.25	0.25
	Pearson	**0.41**	**0.46**
Baseline Train: ru Test: xx	MAE	0.20	0.21
	RMSE	0.26	0.27
	Pearson	*0.32*	*0.33*

Table 1: STL for MT Quality Estimation

3 Transfer of Lexica

Linguistic models can be also transferred through re-using grammatical models trained in a donor language with substitution of the lexicons from a recipient language. For example, a POS tagger can use the transition probabilities from the donor, while the lexical emission probabilities can come from the recipient (**?**; **?**).

Similarly, a traditional MT engine for translation from Ukrainian into English and German can be surpassed by a crude MT pipeline consisting of a direct word-for-word transfer model from Ukrainian into Russian followed by a better resourced model translating from Russian into English and German (**?**). The reason for the success of the pipeline is that the Out-Of-Vocabulary rate is reduced primarily because of the better coverage of the donor lexicon.

Automatic induction of translation lexica between related languages is easier than in the more general case, since in addition to the similarity of the embedding vectors, they often have very similar forms. A reliable lexicon can be produced by combining detection of cognate forms via Levenshtein distance with assessment of semantic similarity via bilingual word embeddings even in the absence of parallel corpora (**?**). One of the problems in transferring the lexica concerns Multi-Word Expressions (MWEs), which tend to differ even for closely related languages. In particular, this concerns fixed-form MWEs without a defined grammatical structure, such as *by and large* or *of course* in English. Such MWEs need to be detected individually in each language and linked to a grammatical model in a donor language via a distributional measure of their similarity to single-word expressions, e.g., *generally* or *definitely* in the examples above (**?**).

In my talk I will also demonstrate an end-to-end example for developing a Named Entity Recognition tagger, which starts with resources available for Slovene and transfers the features derived from a CRF model (**?**; **?**) to other Slavic languages.