

# Cross-Language Text Classification

J. Scott Olsson  
Dept. of Mathematics  
University of Maryland  
College Park, Maryland  
olsson@math.umd.edu

Douglas W. Oard  
College of Information  
Studies/UMIACS  
University of Maryland  
College Park, Maryland  
oard@glue.umd.edu

Jan Hajič  
Institute of Formal  
and Applied Linguistics  
Charles University  
Prague, Czech Republic  
hajic@ufal.mff.cuni.cz

## Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

## General Terms

Algorithms, Design, Experimentation

## Keywords

cross-language text classification, topic classification

## 1. INTRODUCTION

Our goal in cross-language text classification (CLTC) is to use English training data to classify Czech documents (although the concepts presented here are applicable to any language pair). CLTC is an off-line problem, and the authors are unaware of any previous work in this area.

CLTC is motivated by both the non-availability of Czech training data (the case, presently, in our dataset) and the possibility of leveraging different topic distributions in different languages to improve overall classification for information retrieval. Consider, for example, that English speakers tend to contribute more to some topics than their Czech counterparts (e.g., to discuss London more than Prague), so that, having only documents in English, we may expect to do poorly at identifying topics like *Prague*. Czech speakers, on the other hand, often talk about Prague, so that by leveraging Czech data, we might expect to improve on detecting the topic *Prague* in English speakers; and *Prague* in English speakers is exactly the sort of thesaurus label which *information seekers* are most interested in—because it is *rare*. Accordingly, while a lack of Czech training data presently necessitates CLTC, we would have no reason to warrant the method's abandonment if such data were to suddenly become available.

Our dataset is a collection of manually transcribed, spontaneous, conversational speech in English and Czech. English transcripts have human assigned labels from a hierarchical thesaurus of approximately 40,000 labels. Presently, labeled Czech data is not available for classifier training. The hierarchy may be divided into two principle branches, containing 1) concept labels (e.g., *education*) and 2) pre-coordinated place-date labels (e.g., *Germany, 1914 - 1918*).

Copyright is held by the author/owner.  
SIGIR '05, August 15–19, 2005, Salvador, Brazil.  
ACM 1-59593-034-5/05/0008.

## 2. METHOD

A few methods present themselves for CLTC. As we have training data only in English, we may translate all of the Czech data features into English for classification (we refer to this as *English sided* classification). Alternatively, we may translate all English training features into Czech, before classifying in Czech. Finally, we may classify in both directions and combine the evidence. We here confine ourselves to English sided classification, although the concepts may naturally be extended (*mutatis mutandis*) to the Czech and two sided approaches.

Our classification features are vectors of term frequencies in Czech,  $\mathbf{c}$ , and English,  $\mathbf{e}$ .

$$\mathbf{c} = \begin{bmatrix} tf(c_1) \\ tf(c_2) \\ \vdots \\ tf(c_{N_c}) \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} tf(e_1) \\ tf(e_2) \\ \vdots \\ tf(e_{N_e}) \end{bmatrix}$$

A vector's subscript denotes the language from which the term frequencies were *originally drawn* (e.g.,  $\mathbf{e}_e$  denotes a feature vector of English term frequencies that were drawn from an English document). The principle novelty of English sided CLTC then is that, given feature vectors  $\mathbf{e}_e$  and  $\mathbf{c}_c$ , we must produce translated testing vectors,  $\mathbf{e}_c$ , suitable for classification.

The matrix  $E$  represents a probabilistic dictionary mapping between Czech and English terms, such that the  $(i, j)$  element represents the probability that an English word  $e_i$  is the translation of the Czech word  $c_j$ . That is,  $E_{i,j} \equiv P(e_i|c_j)$ ,  $1 \leq i \leq N_e$ ,  $1 \leq j \leq N_c$

$$E = \begin{pmatrix} P(e_1|c_1) & \cdots & P(e_1|c_{N_c}) \\ \vdots & \ddots & \vdots \\ P(e_{N_e}|c_1) & \cdots & P(e_{N_e}|c_{N_c}) \end{pmatrix}$$

By inspection, we see that  $\mathbf{e}_c$  may be reasonably approximated by  $E\mathbf{c}_c \approx \mathbf{e}_c$ , where  $\mathbf{e}_c$  is the left matrix product of the probabilistic dictionary matrix  $E$  and the untranslated Czech feature vector  $\mathbf{c}_c$ . Having attained a set of training vectors  $\mathbf{e}_e$  (via normal indexing) and testing vectors  $\mathbf{e}_c$  (via probabilistic word translation), we are free to continue with classification as before in the monolingual case.

Before documents are indexed, they are parsed and fed into MORPHA [1] and the Czech Feature-Based Tagger [3] for lemmatization. Lemmatization is motivated by both the disparity in morphological richness between English and Czech (which increases the granularity, and thus the noise,

of translation) and the expectation that most of the semantic information associated with words (from which we infer thesaurus labels) is as present in their base forms as it is in their inflections.

The base of the probabilistic dictionary is taken from version 1.0 of the *Prague Czech-English Dependency Treebank* (PCEDT) [4], which contains conditional word-translation probabilities for 46,150 word translation pairs. The dictionary has been derived from a parallel Czech-English corpus based on Reader’s Digest stories, technical texts, and the translation of the Penn Treebank’s WSJ portion into Czech. IBM model 3 has been used in the extraction, and data has been subsequently filtered [2] to avoid most of the noise caused by relatively small datasets.

Indexing proceeds on the English documents by first checking if the term is already present in the probabilistic dictionary. If it is, the term’s frequency is incremented. If the base form for term  $w$  is not present in the dictionary, we hope that the term might be a relevant feature *sans* translation, and therefore augment  $E$  with  $P(e_w|c_w) = 1$  before incrementing  $w$ ’s term frequency. We then index the documents in Czech, although here it is unnecessary to augment the dictionary for previously unseen words (i.e., words not seen in the training documents), as we do not expect to infer a thesaurus label from features never observed in training. The indexed Czech vectors are probabilistically translated via left matrix multiplication of  $E$  and classified using  $k$ NN with symmetric-Okapi. From informal monolingual trials on held out English data, we determined a reasonable choice to be  $k = 20$ .

### 3. EVALUATION

There is currently no labeled Czech data in our dataset. To evaluate our implementation, English sided classification was run on three disjoint segments of 25 Czech sentences each. The segment size was chosen to have roughly 400 words (the average number of words in three minutes of interview). The segments and their ten highest ranked labels were then given to a native Czech speaker for manual relevance assessment. Using the same training set, monolingual English classification was run on four similarly partitioned test segments. The relevance of many labels could not be determined by inspection (e.g., *Poland, 1945* was hypothesized and, while the text made no explicit mention of Poland in 1945, the label was not ruled out). These questionable labelings were all simply assumed to be non-relevant. Table 1 lists precision calculations for both the English sided Czech experiments and monolingual English experiments. Precision was calculated over the five and ten highest ranked thesaurus labels (the complete set) as well as the five highest concept labels alone (that is, without the pre-coordinated place-date labels). Place-date labels may reasonably be excluded from consideration because it is nearly always impossible to assess their relevance to short text segments. On concept labels, the cross-language system performed at 73% of the monolingual precision.

Consider every label assignment to be an independent trial with probability of success  $p$ . Now,  $p$  will vary across thesaurus labels, but the largest  $p$ ,  $p_L$ , will correspond to the label most commonly seen in the training data. If we were to randomly assign any one of the labels to a segment,  $p_L$  would represent an upper bound on the probability of this label being relevant. In this spirit, we can consider  $p_L$  to be

**Table 1: Precision over highest ranked topics**

	top 10		top 5	
	all	all	all	concepts
Czech	.233	.200	.200	.400
English	.450	.325	.325	.550
Czech/English	.518	.571	.571	.727
p-value	$3.4 \times 10^{-6}$	$2.8 \times 10^{-2}$	$2.8 \times 10^{-2}$	$5.0 \times 10^{-6}$
E[prec.]	$p_{LC} = .022$		$p_{LC} .033$	

an upper bound on the probability of success in a series of  $n$  Bernoulli trials, such that an *upper bound* on the chance probability of obtaining  $r$  or more successes in  $n$  trials is

$$P\{r \text{ or more successes}\} \leq \sum_{i=r}^n \binom{n}{i} p_L^i (1 - p_L)^{n-i}. \quad (1)$$

Note that for most  $p$ ,  $p \ll p_L$ , so that we are strongly biasing the test against our method. From inspection, we found  $p_L$  on all labels,  $p_{L_A} = 954/43104$  and  $p_L$  on concepts,  $p_{L_C} = 954/28896$  (both corresponding to the label *extended family members*). The penultimate row of Table 1 lists the p-values calculated for each English sided experiment using Equation 1. We observe that our method is successfully classifying segments across the language barrier. This is likewise confirmed by the final row of Table 1, which lists an upper bound on the expected precision for any of the experiments (an interpretation of  $p_L$ ).

### 4. CONCLUSIONS AND FUTURE WORK

Having introduced the problem of CLTC, we discussed some of its salient features and potential methods for its solution. Our implementation was outlined and preliminary feedback suggests that it is already meeting with some success. Future work will be prompted by the availability of additional testing data, possibly through machine translation of available labeled segments (i.e., to produce labeled pseudo-Czech). This data will allow more extensive evaluation, parameter optimization on held out data, and two sided classifier combination studies.

#### 4.1 Acknowledgments

Thanks to Martin Franz for assisting with relevance assessment. This work has been supported in part by NSF IIS award 0122466 (MALACH) and by the project MŠMT ČR No. MSM0021620838.

### 5. REFERENCES

- [1] G. Minnen et al. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223, 2001.
- [2] J. Curín, M. Čmejrek. Automatic translation lexicon extraction from Czech-English parallel texts. *Prague Bulletin of Mathematical Linguistics*, 71:47–58, 1999.
- [3] J. Hajič. *Disambiguation of Rich Inflection - Computational Morphology of Czech*. Prague Karolinum, Charles University Press, 2004. 334 pp.
- [4] J. Hajič et al. Prague Czech-English Dependency Treebank 1.0. CD-ROM. Catalog no. LDC2004T25. Linguistic Data Consortium, Philadelphia, PA., 2004.