Experiments in Cross-Language Morphological Annotation Transfer

Anna Feldman, Jirka Hana, and Chris Brew

Ohio State University,
Department of Linguistics,
Columbus, OH 43210-1298, USA

Abstract. Annotated corpora are valuable resources for NLP which are often costly to create. We introduce a method for transferring annotation from a morphologically annotated corpus of a source language to a target language. Our approach assumes only that an unannotated text corpus exists for the target language and a simple textbook which describes the basic morphological properties of that language is available. Our paper describes experiments with Polish, Czech, and Russian. However, the method is not tied in any way to these languages. In all the experiments we use the TnT tagger ([3]), a second-order Markov model. Our approach assumes that the information acquired about one language can be used for processing a related language. We have found out that even breathtakingly naive things (such as approximating the Russian transitions by Czech and/or Polish and approximating the Russian emissions by (manually/automatically derived) Czech cognates) can lead to a significant improvement of the tagger's performance.

1 Introduction

Genetically related languages posses a number of properties in common. For example, Czech and Russian are similar in many areas, including lexicon, morphology, and syntax (they have so-called free word-order). This paper explores the resemblances between Czech, Russian, and Polish, as well as exploits linguistic knowledge about these languages for automatic morpho-syntactic annotation without using parallel corpora or bilingual lexicons. Our experiments use these three languages; however, a broader goal of this work is to explore the general possibility of porting linguistic knowledge acquired in one language to another. This portability issue is especially relevant for minority languages with few resources.

Cross-language information transfer is not new; however, most of the existing work relies on parallel corpora (e.g. [7, 11, 12]) which are difficult to find, especially for lesser studied languages, including many Slavic languages. In our work, we explore a new avenue — We use a resource-rich language (e.g. Czech/Polish) to process a resource-poor genetically related language (e.g. Russian) without using a bilingual lexicon or a parallel corpus.

We tag Russian by combining information from a resource-light morphological analyzer ([5]) and information derived from Czech and Polish.

A. Gelbukh (Ed.): CICLing 2006, LNCS 3878, pp. 41–50, 2006.

In the following we report both the overall performance of the model as well as its performance limited to nouns. We deliberately choose nouns, because:

1. As the most open class, nouns are extremely difficult to cover with manually created resources. The set of named entities and proper names is virtually infinite.
2. Nouns is the most challenging category. In the majority of Slavic languages, noun inflection is less systematic than, say, inflection of adjectives or verbs. Moreover, the morphemes are highly homonymous.
3. For practical reasons, we have to limit the scope of our work.

We report tagging accuracy on both the tag as a whole and five categories corresponding to five sub-parts of the complete tag (see Table 1) – part of speech (12 possible values, incl. N/A), detailed part of speech (32, e.g. infinitive or ordinal numeral), gender (5), number (4), and case (8). Note that the number of possible values for detailed part of speech is comparable to the size of Penn Treebank tagset with 36 non-punctuation tags ([8]).

Table 1. Overview and comparison of the tagsets

No.	Description	Abbr.	No. of values		
			Cz	Ru	Po
1	POS	P	12	12	12
2	SubPOS – detailed POS	S	75	32	20
3	Gender	g	11	5	5
4	Number	n	6	4	4
5	Case	c	9	8	9
6	Possessor's Gender	G	5	4	2
7	Possessor's Number	N	3	3	2
8	Person	p	5	5	5
9	Tense	t	5	5	5
10	Degree of comparison	d	4	4	4
11	Negation	a	3	3	3
12	Voice	v	3	3	3
13	Unused		1	1	1
14	Unused		1	1	1
15	Variant, Style	V	10	2	1

2 Tag System

We have adopted the Czech tag system ([4]) for Russian and Polish. Every tag is represented as a string of 15 symbols each corresponding to one morphological category ([6]).

The tagset used for Czech (4290+ tags) is larger than the tagset we use for Russian (about 900 tags). There is a good theoretical reason for this choice – Russian morphological categories usually have fewer values (e.g. 6 cases in

Russian vs. 7 in Czech; Czech often has formal and colloquial variants of the same morpheme); but there is also an immediate practical reason – the Czech tag system is very elaborate and specifically devised to serve multiple needs, while our tagset is designed to capture only the core of Russian morphology, as we need it for our primary purpose of demonstrating portability and feasibility of our technique. The Polish corpus contains 600 tags. This is due to the fact that the original Polish corpus is tagged with a different tagset, which had to be translated into our system and the correspondences are not always isomorphic (see the discussion in section 5.2).

3 Corpora

The experiments below are based on several corpora. One is the first 630K tokens of the morphologically annotated Prague Dependency Treebank ([2]). The other is 630K tokens of the IPI PAN Polish corpus ([9]), translated into our tag system (see 5.2 for the tag translation details).

For development purposes, we selected and morphologically annotated (by hand) a small portion from the Russian translation of Orwell's *1984*. This corpus contains 1858 types (856 types). In the following sections we discuss our experiments and report the results.[1]

4 Morphological Analysis

Our morphological analyzer is a knowledge and labor light system, which takes the middle road between completely unsupervised systems on the one hand, and systems with extensive manually-created resources on the other. Our position is that for the majority of languages and applications neither of these extreme approaches is warranted. The knowledge-free approach lacks precision and the knowledge-intensive approach is usually too costly.

The analyzer is an open and modular system. It allows us to combine modules with different levels of manual input – from a module using a small manually provided lexicon, through a module using a large lexicon automatically acquired from a raw corpus, to a guesser using a list of paradigms, as the only resource provided manually. The general strategy is to run modules that make fewer errors and less overgenerate before modules that make more errors and overgenerate more. This, for example, means that modules with manually created resources are used before modules with resources automatically acquired.

5 Tagging

Our approach assumes that information acquired about a language can be used for processing a related language, in our case information acquired about Czech or Polish can be used to tag Russian.

[1] Note that we do not report the results for tag position 13 and 14, since these positions are unused; and therefore, are always trivially correct.

In ([6]), we describe an n-gram Russian tagger, where transition probabilities were approximated by Czech transition probabilities and emission probabilities were approximated by uniformly distributed output of the morphological analyzer.[2]

In this section, we report on some of the experiments testing both limits and possible enhancements to this basic approach. All the results are summarized in Table 2 (all tokens) and Table 3 (nouns only). In all experiments (except the lower bound), we use the TnT tagger ([3]), which is a second-order Markov model.

5.1 Bounds

Our main practical goal is to develop a portable system for morphological tagging. From the theoretical point of view, we want to understand and isolate general properties of languages that seem to make a difference in the cross-language transfer approach. The experiments discussed in the following two sections simulate two ideal situations: 1) when the word order of a source language is identical to that of the target language (the word order upperbound); 2) when the lexicon of a source language is identical to the target lexicon (the emission upperbound). The upperbounds are given in columns 1 and 2 of Tables 2 and 3.

Table 2. Tagging the Russian Development Corpus: All experiments, all categories

All POS	1	2	3	4	5	6	7	8
Tagger	Max		Even emissions				Cognates	
Accuracy	trans	emis	Cz t	Po t	Interlg	Po&Cz	Manual	Auto
Tags	81.2	95.6	78.6	74.9	79.7	79.1	81.0	80.4
POS	94.5	99.7	92.7	92.0	92.3	91.8	92.8	92.1
SubPOS	87.4	99.7	90.9	90.6	90.0	90.4	91.1	90.5
Gender	93.7	99.7	91.1	90.7	92.1	91.9	92.6	92.2
Number	95.8	99.7	94.0	93.8	94.7	94.6	94.8	94.7
Case	91.7	95.6	87.6	82.6	86.7	86.2	88.3	88.3

Upperbound – Word Order. First, we decided to test how close the Czech and Russian word order is. If they were, it would mean we can train language models relying on word-order, e.g. n-grams, on one language and use it for another.

To measure the upper-bound of the performance of such a model, i.e. the perfect match between the word order in Czech and Russian, we trained the transitions on a small corpus of Russian, and ran [5]'s morphological analyzer to obtain evenly distributed emissions. The results obtained are summarized in column 1 (82.6% accuracy for the nouns). What this means is that the remaining 17.4% deficits are **not** due to word order divergence.

[2] Since Russian and Czech do not use the same words we cannot use the Czech emissions (at least not directly).

Table 3. Tagging the Russian Development Corpus: All experiments, nouns

Nouns	1	2	3	4	5	6	7	8
Tagger	Max		Even emissions				Cognates	
Accuracy	trans	emis	Cz t	Po t	Interlg	Po&Cz	Manual	Auto
Tags	82.6	94.8	65.8	57.0	66.1	66.9	71.3	68.9
POS	97.2	99.7	94.5	93.1	93.1	93.9	94.8	94.2
SubPOS	97.2	99.7	94.5	93.1	93.1	93.9	94.8	94.2
Gender	89.0	98.9	83.5	84.6	84.6	84.3	87.3	85.1
Number	92.0	99.7	90.1	88.4	89.3	89.8	91.2	90.6
Case	87.9	95.6	76.9	65.8	73.8	76.0	79.1	78.2

Upperbound – Lexical Similarities. In the next step, we test how useful the source language lexicon is for the tagging of the target language (here, Russian). We use Czech transitions and Russian emissions, obtained by training TnT on our development corpus. This is the upper-bound performance corresponding to the situation where the source language and the target Russian words behave the same way, all occur in the training data, and we have their perfect translations. The results are in column 2 (94.8% accuracy for nouns). It is clear that the knowledge about Czech-Russian lexical correspondences would definitely help to improve the tagger's performance.

5.2 Approximating Transitions

Below we discuss a number of experiments exploring possibilities of transferring transition probabilities necessary for tagging Russian from a related language.

Approximating by Czech or Polish. In section 5.1, we discuss the word order upperbound. This is an approximation to the performance of the model that would be obtained if there were a perfect correspondence in the word order of Czech and Russian. We wish to know if this result which is obtained by using information about the transitions in the Russian test data, information that we do not have in any realistic situation, can be approximated using Czech.

We train the transitions on 630K Czech tokens, and use the morphological analyzer to create evenly distributed emissions for Russian. The results are given in column 3 in Tables 2 and 3. Such a method approaches the upperbound on transitions.

We also ran an identical experiment with Polish, using the IPI PAN corpus ([9]). This corpus is morphosyntactically annotated, but the structure of its morpho-syntactic tags is different from the tagset we used for Czech and Russian. The repertoire of grammatical categories used in the IPI PAN corpus is different from the Czech tagset. For example, some Polish pronouns are tagged as adjectives, since they have adjectival inflections, whereas the Czech system makes more fine-grained distinctions. Traditional grammatical categories which are represented only partially in the IPI PAN tagset include tense, mood and voice. In addition, since we intentionally did not use a native speaker's expertise for checking the translations (keeping the project resource light), in addition to many differences in the tagset conventions, the translations are not 100% reliable. More importantly, there are obvious linguistic differences between Polish, Czech and Russian. Animacy agreement for adjectives and nouns is obligatory in Polish, whereas in Czech it is manifested only partially and does not exist in Russian at all. There are two types of obligatory copula in Polish (*byc, to*), only one in Czech and none in Russian (for present tense). So, we did not expect a pure Polish model to perform better than the Czech when tagging Russian text.

The results of the experiments are given in Table 2 and Table 3, column 4, for all categories and nouns, respectively. The performance of the Polish model is not as good as of the Czech.

Slavic Interlingua. We discuss one possible solution in detail in [6]. We train the tagger on individual components of the full tag (thus in addition, reducing data sparsity) and then combine them by simple voting. The relative reduction of error rate is 3.3%.

Another possible solution is to create a training corpus which will look more like Russian. Simple "russifications" of Czech lead to 10.5% reduction in relative error rate ([6]).

Every person who knows a Slavic language is able to translate this text, even though it does not belong to any living language: *Korchagin oxvatil glavu rukami i gluboko sa zamyslil. Pred ochami mu prebezhal cely jego zhivot, od detinstva i do poslednix dni. Dobre li on prozhil svoje dvadeset i chetyri let, ili je zle prozhil?*[3] The purpose of this example is to show that it is possible to construct texts that are intelligible to all Slavic speakers.

With a similar idea in mind and with the goal of using minimal resources and minimal knowledge that will not require native speakers' expertise, we decided to create a pseudo-Slavic language which would fuse elements of Czech and Russian and have more Russian-like properties without relying on sophisticated linguistic knowledge. The simple way of doing it is diluting the Czech training data with another resource-rich language which has more Russian-like properties, complementary to Czech. One such language is Polish. We concatenate the Czech 630K tokens with the Polish 630K tokens to create a new training corpus. Polish has some properties that Czech does not. We expect that if we train a tagger on the combination of the two texts, the overall tagging result will improve. The reasons are that negation in Polish is expressed by the particle, whereas in Czech it is expressed by prefixation. Russian is somewhere in the middle – it has cases where negation is a particle, but there is also a class of words, e.g. certain verbs or adverbs, that negate by prefixation. Polish has obligatory genitive of negation. Czech does not have this phenomenon. Russian

[3] This text is a translation of an excerpt from the book *How the Steel Was Tempered* by Nikolai Ostrovsky. Greg Kondrak constructed the translation on the basis of Old-Church-Slavonic, and by consulting translations into the following Slavic languages: Serbo-Croatian, Slovenian, Bulgarian Russian, Ukrainian, Belarusian, High and Low Sorbian, Czech, Slovak, and Polish (from *Introduction to the phonological history of the Slavic languages* by Terence Carlton).

genitive of negation is only partial. With certain noun phrases it is optional, with certain noun phrases it is obligatory. Possessive sentences in Polish are more like Russian (omitting "have", dative constructions) rather than in Czech. The performance of the Russian tagger trained on the Slavic interlingua is given in Table 2, column 5, for all parts-of-speech, and in Table 3, column 5, for the nouns. Our expectations have been met. the interlingua model improves the performance of the Czech model by 1.1% and the pure Polish model by 4.8%, which is a significant improvement. On nouns, The tagging result of the interlingua model is better than the Polish by 9.1%.

Combining Two Language Models. Another possibility is to train the transitions separately on Czech and on Polish and then combine the resulting models into one, taking into account the typological facts about these languages. Based on our linguistic knowledge, we assumed that Polish gender and number for nouns and verbs are more reliable than Czech. The results, given in Tables 2, column 6 for all 12 categories , are better then the models with transition probabilities from individual languages, but not as good as results from the interlingua model. However, in the case of nouns, the situation is reversed. The hybrid model performs better than the interlingua one. The reason, we think, is that the gender and number category is the most relevant for nouns, and our linguistic intuition was correct. We believe that a more sophisticated combination of models would create better results.

5.3 Approximating Emissions – Czech-Russian Cognates

As we said above, since Russian and Czech do not use the same words we cannot use the Czech emissions directly. Instead, the models above approximated Russian emissions by uniformly distributing output of a morphological analyzer. This is a very crude approximation. In this section we explore a different possibility. Although it is true that forms and distributions of Czech and Russian words are not the same, they are also not completely unrelated.

The corresponding Czech and Russian words can be cognates, i.e. historically they descend from the same ancestor root or they are mere translations. We assume that (1) translation/cognate pairs will have similar morphological and distributional properties;[4] (2) cognate words are similar in form.

Manually Selected Cognates. To test the first assumption, we created by hand a list of 202 the most frequent noun Russian-Czech pairs that occurred in our development corpus, which constitutes 60% of all noun tokens in our development corpus. This is clearly not very resource-light, but we do it to

[4] This is obviously an approximation, since certain cognate words in Czech and Russian, even though have similar meanings and morphological properties, do not have the same distributional behavior. For example, the word *ʐivot* means 'belly' in Russian, while *život* means 'life' in Czech; or *krasnyj* means 'red' in Russian, while *krasný* means 'nice' in Czech. Yet, in the former case both words are masculine nouns, and in the latter case both are adjectives.

find out if cognates have any potential.We limit ourselves to nouns due to the reasons outlined above. We used these manual translations for transferring the information about the distribution of Czech words into Russian. In order to do that we normalize and project the tag-frequencies of Czech word into its Russian translation in the case their tags match. The rest of the tags offered by the Russian morphological analyzer for that particular word are redistributed evenly again. For example if $cognate_{czech}$ appears with tag_1 30 times in the Czech corpus, with tag_2 100 times and with tag_3 50 times, after the normalization, the distribution is tag_1 *17*, tag_2 *56*, tag_3 *27*. If the corresponding Russian word is analyzed by the morphological analyzer as either tag_1, or tag_2, tag_4, tag_5, then the new distribution for *ruword* is tag_1 *17*, tag_2 *27*, tag_4 *28*, tag_5 *28*. With this naive procedure, the relative reduction in error rate is 16.1% on nouns, and 11.2% overall – see columns 7, for the detailed information.

Discussion. In our development corpus there are 363 noun tokens, 290 noun types. We are using 202 cognate/translation pairs (types) (= 273 tokens), which means if all these pairs did the expected job, the overall tagging performance on nouns would be (at least) 75.2% (i.e. we would improve the performance on nouns (which is 65.8% without the cognates) by (at least) 9.4%, but in fact we improve only by 5.5%. One of the problems that we have noticed by analyzing the errors is that about half of the Czech manual cognates are not actually found in the 25% most frequent Czech words, which means that we might have been too restrictive by limiting ourselves to the most frequently used words. Nevertheless, even such a naive approach suggests that it is worthwhile to explore this avenue.

Automatic Cognates. In reality, we have no Czech-Russian translations and we do not work with a parallel corpus. In the absence of this knowledge, we automatically identify cognates, using the (normalized by word length) edit distance algorithm. We assume that in a development of a language, vowel changes are more common and less regular than changes of consonants. So, rather than treating all string edits as equal, the operations on vowels have lower costs than on consonants. Yarowsky et al. (2000) use a synchronic version of these assumptions for inflection. This does not require language-intense resources and is general enough to apply to any language we want to work with. In addition, to obtain a more sensitive measure, costs are refined based on phonetic-orthographic regularities, e.g. replacing an ‘h’ with ‘g’ (as in the Czech ‘kniha’ (‘book’) and Russian ‘kniga’ (‘book’) is less costly than replacing ‘m’ with, say ‘sh’. However, we do not want to do a detailed contrastive morpho-phonological analysis, since we want our system to be portable to other languages. So, some facts from a simple grammar reference book should be enough.

Once we identify Czech-Russian cognate pairs automatically, we use the same approach as in the case of manual translations described above. In the case, several cognate candidate pairs have the same edit cost, one of them is selected randomly. Column 8 summarizes the performance of the tagger that uses 149 automatically derived cognates. The cognates we are able to extract by this method help a little in some cases, but the pattern is not as clear as we would like. It looks as if the cognate detector needs further work.

Clearly, the upper-bound on emissions is unreachable, since not for all words in the Russian data there are corresponding Czech words, since even true Russian-Czech cognate pairs might not correspond in their morpho-syntactic behavior. For instance, the words "tema", borrowed from Greek, exists both in Russian and Czech, but in Russian it is feminine, while in Czech it is neuter; moreover, there are definitely false cognates in the two languages, which might mislead the transfer from Czech into Russian (e.g. *matka*: 'uterus' (Russian) vs. 'mother' (Czech)). Finally, our cognate detector is not 100% precise.

6 Discussion and Ongoing Work

This work aims to explore the portability of linguistic knowledge from one language to another. The upper-bounds on TnT transitions and emissions suggest that given we utilize the linguistic knowledge about Czech, Polish and Russian effectively, we can obtain a rather good performance of the tagger. What we showed about Czech, Polish, and Russian surprised us. The model that is trained on a mixture of the two languages, Czech and Polish, outperforms models which were trained on these languages individually. We realize that this is due to the fact that Polish and Czech have complementary Russian-like properties and the mixture of the two creates more Russian-like training data. The fact that the hybrid model outperforms the interlingua model on nouns, where the combination was done using our linguistic intuition about the gender and number assignment in Czech and Polish, is a strong motivation for exploring and exploiting further the linguistic knowledge about the source and the target languages for more accurate tagging.

Our results suggest that the transfer is possible. The system we have developed uses comparable corpora, as opposed to parallel corpora, which makes it very suitable for languages where parallel corpora is not easy to find.

In our ongoing work we are developing an algorithm which will detect cognate stems and generate word forms using the Czech/Russian morphologies. The identification of cognate stems should give more reliable cognate classes, but the next challenge is to map the generated Russian forms into Czech.

Finally, we are extending our work to other languages. We are currently running experiments with Portuguese and Spanish.

Even though the overall performance of our system is not yet comparable to the tagging standard, say, for English, the accuracy of the tagger on the SubPOS position, which is comparable to the Penn Tree bank tagset (32 values) is close to 93%. For many applications this information is useful on its own.

Acknowledgements

We thank Adam Przepiórkowski and Lukasz Debowski for letting us use a fragment of the IPI PAN corpus, as well as for the help with the tagset analysis.

References

1. Agirre E., Atutxa A., Gojenola K., Sarasola K. (2004) Exploring Portability of syntactic information from English to Basque. In Proceedings of LREC 2004, Lisbon, Portugal.
2. Bémová A., Hajič J., Hladká B., Panevová J. (1999). Morphological and Syntactic Tagging of the Prague Dependency Treebank. In Proceedings of ATALA Workshop, Paris, France. pp. 21–29.
3. Brants T. (2000) TnT — A Statistical Part-of-Speech Tagger. Proceedings of ANLP-NAACL. pp. 224–231.
4. Hajic J. (2000) Morphological Tagging: Data vs. Dictionaries. In Proceedings of ANLP-NAACL Conference, pp. 94-101, Seattle, WA, USA.
5. Hana J. (2005) Knowledge and labor light morphological analysis of Czech and Russian. Ms. Linguistic Department. The Ohio State University,
6. Hana J., Feldman A., Brew C. (2004) A Resource-light Approach to Russian Morphology: Tagging Russian using Czech resources. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp.222–229
7. Hwa R., Resnik P., Weinberg A., Cabezas C., Kolak O. (2004) Bootstrapping Parsers via Syntactic Projection across Parallel Texts. Natural Language Engineering 1 (1):1-15.
8. Marcus M., Santorine B., Marcinkiewicz M.A. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics 19(2), 313-330.
9. Przepiórkowski A., (2004). The IPI PAN Corpus: Preliminary version. IPI PAN, Warszawa.
10. Yarowsky D., Wicentowski R. (2000). Minimally Supervised Morphological Analysis by Multimodal Alignment. In Proceedings of the 38th Meeting of the Association for Computational Linguistics. pp. 208-216.
11. Yarowsky D., Ngai G. (2001) Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection Across Aligned Corpora. In Proceedings of NAACL-2001. pp. 200-207.
12. Yarowsky D., Ngai G., Wicentowski R. (2001) Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In Proceedings of HLT 2001, First International Conference on Human Language Technology Research.