Ukrainian part-of-speech tagger for hybrid MT: Rapid induction of morphological disambiguation resources from a closely related language

Bogdan Babych, Serge Sharoff

University of Leeds

b.babych@leeds.ac.uk, s.sharoff@leeds.ac.uk

Abstract: This paper presents a methodology for rapid development of Ukrainian morphological disambiguation resources for a Ukrainian part-of-speech (PoS) tagger and lemmatiser now used in our hybrid MT system. The work is motivated by the need to disambiguate morphological features that result in different translations in rule-based MT and to address out-of-vocabulary (OOV) problem in statistical MT by training factored models. Without morphological disambiguation a larger training or development corpus would be needed to achieve acceptable coverage. Ukrainian, as many other under-resourced languages, does not have publicly released wide-coverage morphological annotation resources in standardised form. However, it has a smaller-scale non-disambiguating tagger with a lexicon of 15k frequent lemmas, which covers 200k unique word forms and generates on average 1.5 ambiguous tags per token (Kotsyba et al., 2009). It is based on a systematic linguistic description and a rich tagset for the Ukrainian morphology developed within the MULTEXT-East project (Erjavec, 2012; Kotsyba et al., 2010). On the other hand, for a better-resourced language, such as Russian, there exist open morphological disambiguation resources, e.g., parameter files for the language-independent TnT tagger trained on a large manually annotated Russian corpus, with estimated tag emission and transition probabilities (Sharoff, Nivre, 2011). Our methodology is based on the assumption that the syntax and morphology in historically related languages change slower than the lexicon, so sentences in them should normally have similar sequences of corresponding morphological features, even when large parts of the lexicon are no longer cognate. Under this assumption, the transition probabilities for the Ukrainian tags are estimated via systematically mapping the tags in the Russian transition parameter file into the Ukrainian tagset. This mapping is not straightforward and requires linguistic expertise in both languages, as even closely related languages have many unique category/value combinations, resulting in different tagsets. Nevertheless, the development time is much smaller than would be required for manually annotating the Ukrainian corpus needed for training the TnT tagger from scratch. Our baseline system described in this paper gives only an unsupervised approximation of the tag sequences in the Ukrainian corpus. It also uses tag emissions that are trivially derived from the seed lexicon, with equal probability settings for tags emitted by ambiguous word forms, and only lemmas mapped or disambiguated from the sample lexicon. However, this baseline is relatively strong as it gives an acceptable accuracy and coverage for morphological annotation tasks. We report evaluation results for the Ukrainian news corpus and we outline techniques for improving the baseline system, which include iterative re-estimation of emission and transition probabilities and iterative learning of rewriting operations for lemmatisation of previously unseen word forms. Resources are made freely available in a public domain on http://corpus.leeds.ac.uk/svitlana/tnt/ua/.[1]

[1] We would like to thank Svitlana Babych (lana.babych@gmail.com) for her contribution to the project – the analysis of the tagset structures and the development of the mapping rule base.

Keywords: PoS tagging; lemmatization; morphological disambiguation; closely-related languages; under-resourced languages; Ukrainian; Russian; Hybrid MT; rapid development

1. Introduction

Creation of morphological analysis and disambiguation tools, especially for highly inflected but under-resourced languages is an important task for MT development, as well as for other natural language processing technologies. In this paper we describe a method for rapid development of resources for Ukrainian morphological disambiguation and present an evaluation of our freely available tagger that uses this methodology. Normally morphological disambiguation tools are trained on disambiguated annotation in a manually checked corpus. Since no such resource is available for Ukrainian, existing taggers leave out the disambiguation stage, only generating a set of all possible tags for each word form (Kotsyba et al., 2009), or do not include the disambiguation by design, e.g., when the intended primary usage is spell checking (Rysin, 2015). Earlier systems used methods of rule-based or semi-supervised disambiguation in the stages of contextual and syntactic analysis (Perebeynos et al., 1989, Gryaznukhina et al, 1999: 51), but no such tools have been released in the public domain, so their accuracy and coverage remains unknown, especially for corpora that include more recent vocabulary.

Our methodology takes an alternative approach: instead of training disambiguation from scratch on a manually checked corpus we rewrite tags for a closely related language (Russian) into the Ukrainian tagset. Russian, as a much better resourced language, has good quality morphological disambiguation resources in standardised formats, used by freely available tagger engines (Sharoff and Nivre, 2011). In our experiment we follow the method used in (Reddy and Sharoff, 2011) by rewriting tags in the parameter file that is used by a language-independent engine of the TnT tagger for calculating tag transition probabilities. The file contains raw frequencies for individual tags in the Russian corpus, and their sequences, up the length of three. The assumption behind this methodology is that morphosyntactic systems in historically related languages change much slower than the lexicon, so such texts should have similar sequences of corresponding morphological features, even when large parts of the lexicon are no longer cognate.

The central problem for our approach is characterising correspondence between non-trivial mismatches in Ukrainian and Russian morphosyntax. Even thought many tags in Ukrainian and Russian have the same configuration of grammatical categories and values, e.g., adjectives in both languages have 7 grammatical values for the case category, 3 for the gender and 2 for the number, but tags often contain information that cannot be mapped in a straightforward way across these two languages, e.g., for Ukrainian – productive synthetic (i.e., one-word) forms for superlative adjectives (*найгарніший* – 'the most beautiful'), synthetic future tense for imperfective verbs (*писатиму* – 'I will be writing'), first-person plural imperative (*йдімо* – 'let's go'), impersonal middle-voice verb forms (*вбито* – 'killed'), more regular use of the vocative case for all Ukrainian nouns (*хлопче* – 'boy!', *чашко* – 'cup!', even though a small number of nouns in Russian have developed new vocative forms: *мам* – 'mum*!*'); for Russian non-mapping features in grammar include active participles (*увидевший* – 'having seen', *плывущий* 'floating'), reflexive participles (*загоревшийся* – 'having started to burn'), short predicative adjectives (*хорош* – 'he is good'). All these forms are

grammatically impossible in the other language. Russian morphological features in tags that are not in the Ukrainian system were rewritten into their functionally closest Ukrainian counterparts, which have similar usage. However, Ukrainian tags missing from the Russian system never appear in the rewritten transition probability file; they only have emission probabilities in the lexicon, and cannot be used for disambiguation of any OOV forms. So rewriting of the Russian tagset in the transition probability file gives only an approximate model of Ukrainian tag combinations.

Our evaluation methodology addresses the question to what extent this approximation would cover disambiguation for a Ukrainian corpus, and how much the mismatches between morphosyntactic systems for this pair of closely related language would interfere with the performance of the tagger.

The paper is organised as follows: Section 2 gives an overview of the use of morphological annotation in MT paradigms and how it affects the requirements for morphological taggers, Section 3 describes the development of the disambiguation resources for the Ukrainian tagger, Section 4 presents tagger evaluation results and the performance of the tagger disambiguation component and Section 5 outlines conclusions and future work. Resources are released in the public domain on http://corpus.leeds.ac.uk/svitlana/tnt/ua/.

2. Use of Morphological resources in MT systems

Morphological processing tools are widely used for a range of computational linguistic tasks, and are often part of a broader processing pipeline, e.g., getting input from text normalisation and feeding into the syntactic and semantic analysis (e.g., Cunningham et al., 2002). These tools work with different linguistic representations and include different processing stages, usually depending on the purpose of the tool. Morphological analysers may or may not include disambiguation, lemmatization or stemming, generation of paradigms, and differ in the level of linguistic details in the tags and forms: some use broad part-of-speech classes (sufficient for less inflected languages), others also process morphological subclasses (regular grammatical categories and their values, such as person, number, gender, case, tense, etc.). MT systems also require specific functionality from the morphological tools, normally, depending on the MT architecture or system type.

If differences between system requirements and the output of morphological processing tools are representational, a new functionality can be added in a straightforward way, but often non-trivial modifications are needed. For example, taggers developed for standard corpus annotation, such as TnT (Brants, 2000) or TreeTagger (Schmid, 1994; 1995) work in the analysis directions, generating morphological tags and lemmas for text forms, however, they cannot be easily extended for working into the generation direction to produce text forms given lemmas and tags – the functionality needed for factored SMT (Koehn, 2010: 316) for combining independently translated lemmas and tags into surface forms (e.g., German lemma Haus + NN.plur → Häuser): in theory it is possible to reverse the direction by tagging and lemmatising a large corpus, but there is no guarantee that it will cover all word forms for all lemmas.

In the statistical MT architecture morphological annotation of corpora is used for training factored models, which allow the system to translate lemmas and morphological features separately and to combine the lexical and morphological factors on the target

side, generating correct inflected target forms even for out-of-vocabulary (OOV) source words, in case if the phrase tables contain translations of their lemmas and morphological features. This addresses the sparse data problem in highly inflected languages, and may potentially affect reordering decisions, checking grammatical coherence and agreement in the target sentences (Kuhn, 2010: 315). Factored models are essential for extending system coverage for language pairs where large parallel corpora are not available. Morphological disambiguation functionality for taggers is used in SMT, primarily for training factored translation and language models on a disambiguated corpus.

In the rule-based MT architecture (RBMT) morphological analysis is a standard processing stage that identifies features of word forms in the source text, such as lemmas (dictionary forms), parts of speech (word classes, e.g., noun, verb, pronoun), additional morphological features, which are used in further stages of syntactic, semantic analysis and bilingual transfer. Correct translation equivalents often rely on successful morphological disambiguation (1):

Their weight **changes.**(VERB.3pers.sing) every day
vs. (1)
Some people record their weight **changes.**(NOUN.plur) every day,

where the word form changes requires different translation equivalents depending on its part of speech). However, RBMT systems traditionally apply rule-based disambiguation techniques, or make an assumption that morphological ambiguity is resolved on higher processing levels, such as the syntactic and semantic analysis (e.g., Odijk, 1993: 33), so their morphological components generated all possible tag+lemma combinations for each word form without the morphology-level statistical disambiguation.

In addition, morphosyntactic representations for RBMT are often more complex and include information needed for highly detailed syntactic analysis and for morphological generation, such as inflection classes, changes in stem, semantic types, and expected morphological values for slots in subcategorization frames. In a hybrid MT framework this information can be partially learnt from large corpora annotated and disambiguated with standard PoS taggers (e.g., Babych et al., 2014).

Our approach to hybrid MT combines a core RBMT system with SMT techniques, exploring synergies between rich linguistic representations and statistical processing methods, which include purpose-built statistical disambiguation modules (Eberle et al., 2012). For example, in SMT target language models can be defined over sequences of any factors or their sets (Kuhn, 2010: 319). We generalise this approach to translation models as well, creating alignments in a richly annotated and morphologically disambiguated corpus across different factors (e.g., alignments between multiword linguistic constructions underspecified either for lexical or for morphological features). Morphological annotation and disambiguation, therefore, is the central component in our research and development of hybrid MT systems, where the challenge is to identify a proper place of statistical and rule-based components within the general architecture, choosing the best performing components from either RBMT or SMT paradigms.

At present, as mentioned in Section 1, publicly available morphological resources for Ukrainian with the large coverage do not include statistical disambiguation component, and this limits their applicability for a number of SMT and Hybrid MT applications. Our approach addresses this problem by deriving disambiguation resources for Ukrainian from a better-resourced closely related language.

3. Development of the morphological disambiguation resources for Ukrainian

3.1. The overview of the tagger

We developed morphological disambiguation resource for Ukrainian, in a standardised format of tag transition frequencies file for the language-independent engine of the TnT tagger (Brants, 2000). In the first stage the morphological lexicon of ~15k lemmas (~200k inflected forms) from the Ukrainian non-disambiguating tagger (Kotsyba et al., 2009) has also been converted into the format used by the TnT tagger, into representation of the tag emission frequency file.

The lexicon contains only frequent Ukrainian words (c.f. commercial wide-coverage systems for Ukrainian use over 100k lemmas). However, this lexicon covers about 93% of tokens in Ukrainian news texts (~90% excluding digits and punctuation). The TnT tagger generates tags for missing words using the tag transition frequencies, as we will explain below, but lemmatization is currently available for the word forms from this lexicon. An alternative solution is to use a much larger Ukrainian lexicon developed for open-source Ukrainian spelling platforms, such as ispel-uk (Rysin, 2015). However, the advantage of Kotsyba et al.'s Ukrainian morphological lexicon is that the tagset has been developed in the standardised MULTEXT format (Erjavec, 2012; Kotsyba et al., 2010), which makes the mapping much easier between tagsets of the closely related languages. It also allows us to test the performance of our disambiguation more clearly on the larger number of word forms missing from the tag emission lexicon. Our future work will include integration of Rysin's and Kotsyba et al.'s lexicons, to improve tagging accuracy and lemmatization coverage. Table 1 describes the size and tag distribution in Kotsyba et al.'s lexicon.

Unique lemmas	15,162
Unique {word forms+pos tag} combinations	300,292
Unique word forms	205,348
unique tags (pos+morphology)	1,239
Average word-form ambiguity (tags per word form)	1.46
Average paradigm size (word forms per lemma)	13.54

Table 1. Ukrainian lexicon from (Kotsyba et al., 2009) used for tag emission file

Emission frequencies are all set to the default value of *"1"*, because disambiguated tag frequencies in the Ukrainian corpus is unknown. This file looks as shown in Figure 2.

```
сльоза  1   Ncfsnn  1
сльозам 1   Ncfpdn  1
сльозами    1   Ncfpin  1
сльозах 1   Ncfpln  1
сльози  4   Ncfsgn  1   Ncfpvn  1   Ncfpnn  1   Ncfpan  1
сльозо  1   Ncfsvn  1
сльозою 1   Ncfsin  1
сльозу  1   Ncfsan  1
сльозі  2   Ncfsln  1   Ncfsdn  1
```

Figure 2. Tag emission file

In this example some inflected word forms of the Ukrainian noun *сльоза* (‘*sl’oza*’ – ‘a teardrop’) are listed with their default emission frequencies. All belong to the part of speech noun, but differ in their values of the grammatical categories of Case and Number. The form *сльозі* (‘*sl’ozi*’, in the last line) is ambiguous between {Number.Singular, Case.Locative} and {Number.Singular, Case Dative} (‘in a *teardrop’* vs. ‘*to a teardrop’)*; a more complex ambiguity exists for the form *сльози* (‘*sl’ozy*’, in the line 5), which in the spoken form either has the stress on the first syllable, which is ambiguous between {Number.Plural, Case.Nominative | Case.Accusative | Case.Vocative} (a systematic ambiguity for all Ukrainian inanimate plural nouns); or it has the stress on the second syllable, having the values of {Number.Singular, Case.Genitive}.

As stress is not marked in writing, all four possibilities are added to the list of ambiguous tags. In the general case it is not possible to estimate if any of the {Number,Case} combinations would be more frequent in corpus: this depends on a specific lexical item. For example, the same stress-related ambiguity between {Number.Plural, Case.Nominative} and {Number.Singular, Case.Genitive} (sl’***o”***zy – sl’oz***y”***) applies for a number of other nouns. In a 500k corpus of the Ukrainian fiction prose, which has been manually disambiguated for the frequency dictionary of the 20-th century Ukrainian prose (Perebyinis, (Ed.), 1984) the plural form is normally more frequent for nouns which denote objects existing in pairs, e.g.: ‘hands’, ‘feet’ (r**u”**ky, n**o”**hy), but singular forms are more frequent for nouns that exist as single objects, e.g., ‘head’ (holov**y”**). For this reason all the frequencies in this tag emission file have been set to the same default value, which might cause a certain number of errors, but allows us to have a working system without the need to manually annotate a large Ukrainian corpus.

In our implementation, the tag probabilities and sequence probabilities are estimated from the transition frequency file. The TnT engine uses this file for morphological disambiguation, so rapid induction of this information for the Ukrainian tagset allows to create the missing morphological disambiguation tools for Ukrainian; so it is the main purpose of our experiment. The transition frequency file contains corpus frequencies for single tags, and for tag sequences of two and three tags. The example of the data in this file is given in Figure 3.

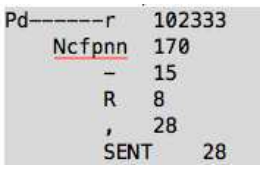

Figure 3. Transition frequencies for tags

The tags sequences here are represented in tabulated format (frequencies for each tag after the tab show the number of occurrence with the preceding higher-level tags). Normally frequency counts for this file should be calculated from a manually checked tagged and morphologically disambiguated corpus. Such corpus needs to be representative, covering a reasonable number of potential tags and ambiguous word forms, which have different tags in different morphosyntactic contexts. Importantly, frequency counts for tags sequences need to be large enough to converge on their true probabilities (when absolute counts are converted into relative frequencies, i.e., divided by the length of the corpus). Understandably, creation of such a corpus for a new under-

resourced language such as Ukrainian would be a large-scale time-consuming task, which would involve an extensive manual annotation effort, often not feasible for teams, who develop freely available morphological resources.

Our alternative approach described further in this section involves a much smaller, quicker, but a more linguistically qualified effort for induction of transition frequency information for Ukrainian from tag transition frequencies calculated on a Russian morphologically disambiguated corpus. Russian is a much better resourced language, with a high quality manually disambiguated corpus, developed within the Russian National Corpus (RNC) project (Sharoff, 2005).

Our assumption is that in a closely related language the ordering of parts of speech, their morphological categories and values will be similar, mainly because the grammar system in languages undergoes historical changes at a much slower rate compared to the lexical or phonological systems. As a result, a much higher proportion of morphosyntactic similarities between Ukrainian and Russian exist compared to a relatively smaller amount of similarities in the lexicon: grammatical similarities include the system of grammatical categories, the inflection and declension systems, large parts of the verb morphology, the word order), even though the Ukrainian and Russian languages separated by the end of the Early Proto-Slavonic period, around the 7th century AD, acquiring their distinctive phonological, grammatical and lexical features and integrating the elements of the substratum languages: Iranian substratum for Ukrainian (e.g., fricative h ← g), and Baltic for Russian (e.g., a ← o in unstressed position) (Pivtorak, 1988: 52, 92; Schenker, 1993: 114).

We suggest that a way for rapid development of morphosyntactic disambiguation resources for Ukrainian is to reuse frequencies of tag transitions calculated on a disambiguated corpus for a closely related language, by translating the Russian tags set into corresponding Ukrainian tags. This development route is much more feasible for smaller but highly qualified research teams, who work on similar projects of creating morphologically disambiguation tools for other under-resourced languages, and can be replicated for this scenario.

A challenge for our approach is that the tag sets for closely related languages are not the same, so Russian tags have to be rewritten into their closest morphologically equivalent tags for Ukrainian. The task rewiring tagsets becomes (to some extent) similar to the development of a rule-based MT system, with similar imperfections and approximate results, mainly because of systematic morphological differences between the languages. Categories and values within tags are often structured differently: as discussed in Section 1, some grammatical values for categories could be missing in the Ukrainian target system (e.g., reflexive participles, short predicative adjectives), some are missing in the Russian source system (e.g., vocative case, superlative adjectives, imperative first person plural, impersonal and imperfective future verbs, etc.), which would necessarily lead to approximations and imperfect technical decisions in finding and mapping the corresponding tags.

Another theoretical limitation for our route is that there are contrastive distributional and usage differences between Ukrainian and Russian for certain parts of speech. For example, Ukrainian would use a different syntactic perspective for the example in (4), suggested by Gryaznukhina, as the reflexive participle with its corresponding structural links cannot be used in the Ukrainian sentence:

Ru:	*Дети,*	*увидевшие*	*загоревшуюся*	*крышу,*	*закричали*	
	N	PastParticpl	PastParticiplRefl	N	V	
	Children, seeing		burning	roof,	shouted	
	‘Children who saw the burning roof, shouted’					
Uk:	*Побачивши,*	*що*	*загорівся*	*дах,*	*діти*	*закричали* (4)
	V.Adv.Past	Conj	V	N	N	V
	Having seen ,	that	burns	roof,	children	shouted
	‘Having seen that the roof burns, children shouted’					

If the differences highlighted by this example are frequent, the performance of our morphological disambiguation method will be affected, as this would undermine the assumption about strong parallelism between tag sequence probabilities in Ukrainian and Russian. On the other hand, a working disambiguation system for Ukrainian developed with our methodology allows us to empirically test to what extent such sequence asymmetries are widespread and how much they affect the performance in practice (e.g., for specific subject domains and genres).

In the final processing stage we lemmatize word forms using their disambiguated tags and the mapping {word form + tag} → lemma, which we derive from Kotsyba et al.’s lexicon. In most cases the combination of a word form and its PoS tag unambiguously determines lemma, so this mapping is deterministic. While tags are generated both for known and unknown words (using tag transition probabilities), lemmatization at the moment only covers 15k frequent Ukrainian lemmas in the lexicon. However, this problem will be addressed in future using a larger freely available lexicon and heuristic lemmatization procedures, learnt from the existing lemmatized lexicon, for rewritings inflected word forms.

The tagger produces a standard word-per-line tab-separated format with three fields for the form, POS tag and lemma, as used in IMS Corpus Workbench (Evert, 2010), see Figure 5:

```
Які     Pq----pna       який
там     Pd------r       там
сльози  Ncfpnn  сльоза
!       SENT    !
Хмари   Ncfsgn  хмара
лебедині        Ncfsdn  <unknown>
плили   Vmpis-p плисти
над     Spsa    над
нами    Pp-1-ypin       ми
,       -       ,
струмувала      Vmpis-sf        <unknown>
даль    Ncfsnn  <unknown>
,       -       ,
і       Q       і
вірилось        Vmpis-sn        віритися
,       -       ,
що      Q       що
лиш     Q       лиш
на      Q       на
те      Q       те
людині  Ncfsdy  людина
,       -       ,
щоб     Q       щоб
радощі  Nc-pan  радощі
підкреслити     Vmen    підкреслити
,       -       ,
печаль  Ncfsnn  печаль
.       SENT    .
```

Figure 5. Output generated by the Ukrainian tagger

3.2. Mapping of Russian tags into the Ukrainian tagset

Both the Ukrainian and Russian tagsets used in our experiment have been developed within the MULTEXT-East project, which facilitates rewriting task, as common symbols are typically used for the same values of grammatical features. However, the order of these features and the structure of the tags (i.e., the set of features they include) are often different, which normally reflects differences in morphological systems.

	Russian	Ukrainian
1	^Nc$	Nccsnn
2	^Nc\-p([ngdail])$	Nc-p\g<1>n
3	^Npc(s|p)([ngdail])y$	Npm\g<1>\g<2>y
4	^Mo\-pa$	Mlo-pan
5	^Momsa$	Mlomsay
6	^M([coml])([mfn\-])([\-])([ngdail])$	Ml\g<1>\g<3>p\g<4>
7	^M([coml])([mfn\-])([sp\-])([ngdail])$	Ml\g<1>\g<2>\g<3>\g<4>
8	^Mo[mfn][sp][ngdail]----n$	Mr
9	^Mo$	Mr
10	^Mc$	Mlc-pn
11	^Mc-s$	Mlc-pn
12	^Sp\-([gdail])$	Sps\g<1>
13	^Spc$	Spcg
14	^P\-\-\-pa$	Pd--n-sga
15	^P\-\-\-sg$	Pd---ypaa
16	^P\-\-msa$	Pd--mnsaa
17	^P([pdisqrxz\-])([12])([mfn\-])([sp])([ngdail])$	Pp-\g<2>\g<3>y\g<4>\g<5>n
18	^P([pdisqrxz\-])([3])([mfn\-])([sp])([ngdail])$	Pp-\g<2>\g<3>-\g<4>\g<5>n
19	^P([pdisqrxz\-])([\-])([mfn\-])([sp])([ngdail])$	Pd-\g<2>\g<3>-\g<4>\g<5>a
20	^P\-\-([\-])([p])([n])-([y])$	Pz--\g<1>-\g<2>\g<3>a
21	^P\-\-([mnf\-])([sp])([ngdail])-([yn])$	Pz--\g<1>\g<4>\g<2>\g<3>n
22	^P\-\-\-\-([gdali])$	Px---y-\g<1>n
23	^P$	Pd------r
24	^Vmi\-\-sn([am])\-([pe])$	Vm\g<2>is-sn
25	^Vmi\-([12])([sp])\-([am])\-p$	Vmpip\g<1>\g<2>
26	^Vmi\-([12])([sp])\-([am])\-e$	Vmeif\g<1>\g<2>
27	^Vmg----[am]-p$	Vmpgp
28	^Vmg[fs]---[am]-e$	Vmeg
29	^Vm\-([pf])([123])([sp])\-([am])\-([pe])$	Vm\g<5>i\g<1>\g<2>\g<3>
30	^Vm\-([s])\-([s])([mfn])([am])\-([pe])$	Vm\g<5>i\g<1>-\g<2>\g<3>
31	^Vmi([ps])([13])pmps([p])$	Vmpip\g<2>p
32	^Vmi([ps])([13])pmps([e])$	Vmeif\g<2>p
33	^V([ma\-])([imng])([pfs\-])([123\-])([sp\-])([mfn\-])([apm\-])([\-])([peb\-])$	V\g<1>\g<9>\g<2>\g<3>\g<4>\g<5>\g<6>
34	^Vmp([ps])\-([sp])([mfn\-])([a])([fs])([pb])([ngdali])$	Ap-\g<3>\g<2>\g<7>\g<5>-\g<6>\g<4>\g<1>
35	^Vmp([ps])\-([sp])([mfn\-])([m])([f])([pbe])([ngdali])$	Ap-\g<3>\g<2>\g<7>\g<5>-\g<6>p
36	^Vmp([ps])\-([sp])([mfn\-])([p])([s])([pbe])$	Ap-\g<3>\g<2>nf-\g<6>\g<4>
37	^Vmp([ps])\-([sp])([mfn\-])([ap])([fs])([pbe])([ngdali])$	Ap-\g<3>\g<2>\g<7>\g<5>-\g<6>\g<4>
38	^Vm\-([ps])\-pn([a])f([e])i$	Ap--pif-\g<3>a
39	^Vm\-([ps])\-pn([a])f([pb])i$	Ap--pif-\g<3>a\g<1>
40	^Vm\-([ps])\-pn([pm])f([peb])i$	Ap--pif-\g<3>p
41	^Vm\-([ps])\-([sp])([mfn\-])([a])([fs])([pb])([ngdali])$	Ap-\g<3>\g<2>\g<7>\g<5>-\g<6>\g<4>\g<1>
42	^Vm\-([ps])\-([sp])([mfn\-])([m])([f])([pbe])([ngdali])$	Ap-\g<3>\g<2>\g<7>\g<5>-\g<6>p
43	^Vm\-([ps])\-([sp])([mfn\-])([p])([s])([pbe])$	Ap-\g<3>\g<2>nf-\g<6>\g<4>
44	^Vm\-([ps])\-([sp])([mfn\-])([ap])([fs])([pbe])([ngdali])$	Ap-\g<3>\g<2>\g<7>\g<5>-\g<6>\g<4>
45	^Vm\-([ps])\-([sp])([mfn\-])([a])([f\-])([p])$	Ap-\g<3>\g<2>nf-\g<6>\g<4>\g<1>
46	^Vm\-([ps])\-([sp])([mfn\-])([a])([f\-])([be])$	Ap-\g<3>\g<2>nf-\g<6>\g<4>
47	^Vm\-([ps])\-([sp])([mfn\-])([m])([f\-])([pbe])$	Ap-\g<3>\g<2>nf-\g<6>p
48	^Afp([m\-])([sp])af?$	Afp\g<1>\g<2>afn
49	^Afp([m])([s])\-([s])$	Afp\g<1>\g<2>nf
50	^Afp([mfn\-])([sp])\-([sf])$	Afp\g<1>\g<2>n\g<3>
51	^Afpnpif$	Afp-pif
52	^A([fs])([pcs])([mfn\-])([sp])([nagdli])$	A\g<1>\g<2>\g<3>\g<4>\g<5>f
53	^Af([pc])$	Af\g<1>msnf
54	^Af([pc])\-\-\-f$	Af\g<1>msnf
55	^Vmi$	Vmbn
56	^Vmi\-\-\-\-m$	Vmbn
57	^Vm$	Vmbn
58	^Vm\-\-\-\-\-\-\-\-d$	Vmbn
59	^Vm------f$	Vmbn
60	^Vm-----mf$	Vmbn
61	^Vm---sn---d$	Vmbn
62	^Vm--1p---p$	Vmbn

Table 2. Regular expressions for mapping Russian tags into Ukrainian tagset

Tag rewriting is performed with 62 regular expression mappings shown in Table 2. It can be seen from the Table that the main effort went into rewriting of Russian tags for verbs, which is understandable, given major differences between Ukrainian and Russian verb forms systems, mainly the structure of participles. For example, many participle forms belong to the verbal paradigm in Russian, but in Ukrainian they function as adjectives, which is reflected in their annotation, covered by the regular expressions 34 through 47. Other operations included rewriting Russian predicative adjectives, inserting some missing categories (e.g., animate for adjectives in the accusative case), or changing order of grammatical categories in tags. However creation of these rewriting rules can be done much faster than annotating and manually checking a representative sample of a Ukrainian corpus for training the disambiguation tools for the tagger.

4. Evaluation of the Ukrainian tagger

Evaluation of the tagger performance is done on a corpus of Ukrainian news texts, on a section of about 1000 words selected from the 250MW corpus. In this evaluation experiment we identified only broader part-of-speech errors (which is mostly needed for lemmatisation), but not errors within morphological subclasses (categories and values). Performance parameters of the tagger are shown in Table 3.

	Word forms + punctuation	**(Percent)**	**Excluding punctuation**	**(Percent)**
Sample length	999	100.0%	793	100%
Punctuation (+numerals)	206	20.6%		
Unknown words	74	7.4%	74	9.3%
Known words (coverage)	**925**	**92.6%**	**719**	**90.7%**
Wrong PoS	54	5.4%	54	6.8%
Correct PoS (performance)	**945**	**94.6%**	**739**	**93.2%**
				Accuracy for Unknown:
Correct Unknown	**53**			**71.6%**
Errors:	Error type counts:	Error type percentages:		Errors for Unknown
Wrong Unknown	21	38.9%		28.4%
Disambiguation errors	23	42.6%		
Lexicon errors	10	18.5%		
(All error types)	(54)	(100%)		

Table 3. Evaluation of the Ukrainian TnT tagger

It can be seen from Table 3 that the tagger achieves over 90% lexicon coverage, with correct tags generated overall for 93% word forms in the corpus (known and unknown words put together). Among the 74 unknown word forms, 71.6% are tagged correctly, and 28.4% have wrong tags.

Also, 54 found errors can be classified into 3 types: Wrong tag for unknown word forms (38.9% of all errors), disambiguation errors for known forms (42.6%) and the errors coming from wrong annotation of the lexicon (18.5%).

For our tasks performance of the Ukrainian TnT tagger (93.2%) is acceptable (given the small size of the lexicon), its performance on unknown words, where the system tries to guess a tag using our transition frequency file, is also relatively high (71.6%).

5. Conclusions and future work

In our approach to rapid development of morphological disambiguation tools for an under-resourced language (Ukrainian) the tag transition probabilities are deduced from a table of frequencies that is calculated on a manually checked corpus of Russian, which is a better-resourced closely related language. The results indicate that our approach has potential, as it requires a smaller but a more qualified development effort, as it involves non-trivial rewriting of tags that needs to reflect differences in morphological system between the two languages.

Future work will include a more systematic evaluation of the tagger performance on different text types and using a finer grained evaluation of morphological classes and sub-classes (parts of speech and the values of grammatical categories), improving the coverage of the lexicon, lemmatization and the accuracy of tagging via iterative re-estimation of the tag emission and transition probabilities, combining statistical and rule-based disambiguation techniques and learning token rewriting operations for lemmatization from the examples in the lexicon.

Bibliography

Babych, B., Geiger, J., Rosell, M. G., and Eberle, K. (2014, April). Deriving de/het gender classification for Dutch nouns for rule-based MT generation tasks. In Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra)@ EACL (pp. 75-81).

Brants, T. (2000, April). TnT: a statistical part-of-speech tagger. In Proceedings of the sixth conference on Applied natural language processing (pp. 224-231). Association for Computational Linguistics.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. In Proc. 40th Anniversary Meeting of the Association for Computational Linguistics (ACL).

Eberle, K., Geiß, J., Ginestí-Rosell, M., Babych, B., Hartley, A., Rapp, R., Sharoff, S and Thomas, M. (2012, April). Design of a hybrid high quality machine translation system. In Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra) (pp. 101-112). Association for Computational Linguistics.

Evert, S. (2010). The IMS Open Corpus Workbench Corpus Encoding Tutorial, http://cwb.sourceforge.net/files/CWB_Encoding_Tutorial.pdf

Erjavec, T. (2012): MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. Language Resources and Evaluation, 46/1, pp. 131-142.

Gryaznukhina, T.A., (Ed.) (1999). Syntactic analysis of scientific texts on computers. (Sintaksicheskiy analiz nauchnogo teksta na EVM) Naukova Dumka, Kiev. (In Russian).

Koehn, P. (2009). Statistical machine translation. Cambridge University Press. P. 314_

Kotsyba, N., Mykulyak, A., and Shevchenko, I. V. (2009). UGTag: morphological analyzer and tagger for the Ukrainian language. In Proceedings of the international conference Practical Applications in Language and Computers (PALC 2009). Software URL: Description URL:

Kotsyba, N., Shevchenko, I., Derzhanski, I. and Mykulyak, A. (2010) MULTEXT-East Morphosyntactic Specifications, Version 4. 3.11. Ukrainian Specifications: URL: http://nl.ijs.si/ME/V4/msd/html/msd-uk.html. Accessed on 28/04/2016

Odijk, J. E. J. M. (1993). Compositionality and syntactic generalizations.

Perebeynos, V. I., Darchuk, N.P. and Gryaznukhina T.A. (1989). Morphological analysis of scientific texts on computers. (Morfologicheskiy analiz nauchnogo teksta na EVM) Naukova Dumka, Kiev. (in Russian)

Perebyinis, V.S. (Ed.) (1984). Frequency dictionary of the modern Ukrainian fiction prose (Chastotnyi slovnyk suchasnoyi ukrayins'koyi khudozhn'oyi prozy). Naukova Dumka, Kyiv. (In Ukrainian).

Pivtorak, H. P. (1988). Forming and dialectal differentiation of the old Ukrainian language. (Formuvannya i dialektna dyferentsiatsiya davn'orus'koyi movy). Naukova Dumka, Kyiv. (in Ukrainian).

Reddy S. and Sharoff, S. Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources. In Proc. 5th International Joint Conference on Natural Language Processing, Chiang Mai, 2011. http://www.mt-archive.info/IJCNLP-2011-Reddy.pdf

Rysin, A. (2015). Project to generate POS tag dictionary for Ukrainian language. URL: https://github.com/arysin/dict_uk. Accessed on 28/04/2016.

Schenker, A. M. (1993). Proto-Slavonic. In: Comrie, B & Corbett, G. (Eds.). The Slavonic Languages. Routledge: London, New York. Pp. 60-121.

Schmid, H. (1994, September). Probabilistic part-of-speech tagging using decision trees. In Proceedings of the international conference on new methods in language processing (Vol. 12, pp. 44-49).

Schmid, H. (1995). Treetagger: a language independent part-of-speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 43, 28.

Sharoff, S. (2005) Methods and tools for development of the Russian Reference Corpus. In D. Archer, A. Wilson, and P. Rayson, editors, Corpus Linguistics Around the World, pages 167–180. Rodopi, Amsterdam.

Sharoff, S. and Nivre, J. (2011). The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In Proc. Dialogue, Russian International Conference on Computational Linguistics, Bekasovo.