Unsupervised induction of Ukrainian morphological paradigms for the new lexicon: extending coverage for Named Entities and neologisms using inflection tables and unannotated corpora

Bogdan Babych
Centre for Translation Studies
University of Leeds
UK
b.babych@leeds.ac.uk

Abstract

The paper presents an unsupervised method for quickly extending a Ukrainian lexicon by generating paradigms and morphological feature structures for new proper names and neologisms, which are not covered by existing static morphological resources. This approach addresses a practical problem of modelling paradigms for entities created by the dynamic processes in the lexicon: this problem is especially serious for highly-inflected languages in domains with specialised or quickly changing lexicon. The method uses an unannotated Ukrainian corpus and a small fixed set of inflection tables, which can be found in traditional grammar textbooks. The advantage of the proposed approach is that updating the morphological lexicon does not require training or linguistic annotation, allowing fast knowledge-light extension of an existing static lexicon to improve morphological coverage on a specific corpus. The method is implemented in an open-source package on a GitHub repository https://github.com/bogdanbabych/paralex4morphosyntax. It can be applied to other low-resourced inflectional languages which have internet corpora and linguistic descriptions of their inflection system, following the example of inflection tables for Ukrainian. Evaluation results show consistent improvements in coverage for Ukrainian corpora of different corpus types.

1 Introduction

> "Our language can be regarded as an ancient city: a maze of little streets and squares, of old and new houses, of houses with extensions from various periods, and all this surrounded by a multitude of new suburbs with straight and regular streets and uniform houses." (Wittgenstein, 2009)

This metaphor from Wittgenstein's 'Philosophical Investigations' may be applied to two aspects of the natural language lexicon, which so far have received little attention in computational linguistics. Firstly, like a city, the lexicon constantly evolves, reflecting political and technical changes in the society that take place very rapidly, so it may be insufficient to design static lexical resources and to expect that they would give the same high level of corpus coverage once and for all: the lexicon needs to be constantly updated to reflect live changes in the system. Secondly, even though there may be many irregularities in the lexicon, similar to 'a maze of little streets', this more often happens with an older lexical core, while new words typically follow more 'straight and regular' patterns, so the task of updating the lexicon for natural language applications may be facilitated by this tendency.

This paper investigates the extent of the new lexicon problem for different types of Ukrainian corpora and further proposes and evaluates a knowledge-light approach to extending lexical coverage of morphological resources to neologisms (new words, meanings or usages) and new single-word Named Entities (proper names) which follow regular inflectional patterns.

Morphological annotation of the lexicon is an important component for many natural language processing pipelines, such as part-of-speech tagging, morphological disambiguation, parsing, semantic analysis, as well as for applications such as machine translation, information extraction, terminology detection, etc. For example, in part-of-speech tagging the morphological lexicon normally supplies lemmas and associated sets of possible parts-of-speech and values of morphological categories for each token (e.g., for Ukrainian this would be values for the grammatical case: nominative, genitive, dative, accusative, instrumental,

locative, vocative; number: singular, plural; gender: masculine, feminine, neuter; person: 1st, 2nd, 3rd; mood: indicative, imperative, subjunctive; tense: past, present, future, etc.). The tagger then resolves any potential ambiguity using transition probabilities, trained neural networks, etc. For any language the creation of a morphological lexicon is difficult, because of a large number of lexical types needed to achieve good corpus coverage and also because of irregularities in word paradigms (systems of inflected word forms, lemmas and associated morphological features). For highly inflected Slavonic languages the creation of the morphological lexicon is even more challenging, since most words have complex morphological paradigms, which require fine-grained annotation of parts-of-speech and their grammatical subcategories. Creation of high-quality morphological resources for these languages often requires an extensive effort over many years.

Like the majority of other Slavonic languages, Ukrainian is a highly inflected language with the 'synthetic' grammar structure (where grammatical relations are predominantly marked within content word forms), so the task of morphological paradigm generation for it is not trivial. It is also more critical for the accuracy of related tasks, such as part-of-speech tagging, because of a larger potential number of combinations of possible morphological values: it is harder to guess the correct part-of-speech tag based on neighbouring tags in the case of a missing word form. Ukrainian paradigms for inflected parts of speech have between 7 and 28 distinct morphological feature combinations and associated word forms for a single lemma, and there is both regular and irregular ambiguity within and across different parts-of-speech and lexicogrammatical classes of words (i.e., animate vs. inanimate nouns, perfective vs. imperfective verbs). In Ukrainian, as in other highly-inflected languages most morphological information is supplied within the word rather than by the context, so lexical gaps are more detrimental for correct prediction of inflected word forms and their morphological characteristics.

For Ukrainian there exist wide-coverage lexical resources (see Section 2), however, extending them in a traditional rule-based way would involve continuous annotation effort requiring linguistic expertise and near-native knowledge of the language, making it hard to keep up with most recent lexical developments.

The approach proposed in this paper is designed for the scenario where for a highly-inflected language there exists a hand-crafted static morphological lexicon that covers potentially irregular and more frequent lexical core. For extending this lexicon to cover new regularly inflected entities I use an internet corpus and small inflection tables from grammar textbooks, e.g., (Hryshchenko et al., 1997), (Press and Pugh, 2015): such resources would often be available for other low-resourced languages, since the tasks that would require linguistic expertise (i.e., creating the core lexicon and inflection tables) need to be done only once, so paradigms for new entities can be automatically created whenever a new corpus becomes available. Core static morphological lexicons have been developed for several low-resourced languages, either as stand-alone resources or within shared frameworks, such as Universal Dependencies (Nivre et al., 2016), Apertium (Forcada et al., 2011) (in the context of Machine Translation) or Grammatical Framework (Ranta, 2011) (in limited subject domains). However, the task of extending morphological lexicon in response to dynamic processes in the lexical system, emergence of neologisms, new terminology or Named Entities has not been systematically addressed so far.

The paper is organised as follows. In Section 2 I review some of the previous work in the area, in Section 3 I describe the algorithm, datasets and an experiment on generating paradigms, in Section 4 I present experimental results on comparative evaluation of lexical coverage on different corpora for the baseline static morphological lexicon and for the extended paradigms which cover the new lexicon. Section 5 presents a discussion of examples of identified new entities and in Section 6 I summarise conclusion and ideas for future work.

2 Previous work

Several projects have addressed the problem of developing the Ukrainian morphological lexicon and morphological disambiguation strategies: (Perebyjnis et al., 1989), (Gryaznukhina, 1999), (Rysin and Starko, 2019), (Kotsyba et al., 2009), (Kotsyba et al., 2010), (Babych and Sharoff, 2016). For the experiments presented in this paper I use the most complete morphological toolkit from (Rysin and Starko, 2019), which in its current implementation contains a wide-coverage lexicon:

Corpus	No of words	No of sent
News	461,451,019	31,021,650
Wikipedia	185,645,357	15,786,948
Fiction	18,323,509	1,811,548
Law	578,988,264	29,208,302
Total	1,244,408,149	77,828,448

Table 1: Description of Ukrainian corpora from (Dyomkin et al., 2019).

366,846 Ukrainian lemmas, which are expanded into 5,690,688 word forms with corresponding morphological feature combinations.

We evaluate the coverage of this lexicon on large Ukrainian corpora collected in lang-uk project (Dyomkin et al., 2019), Table 1 is taken from this source, which describes these collections.

Detailed overviews of different approaches to developing morphological lexicons can be found in (Ahlberg et al., 2015), (Koskenniemi et al., 2018) and (Fam and Lepage, 2018). For our purposes the existing approaches can be characterised by their application scenarios and assumptions about available datasets. Interesting work has been done within the neural, supervised and semi-supervised frameworks, e.g., (Ahlberg et al., 2015), (Ahlberg et al., 2014), (Koskenniemi et al., 2018), (Silfverberg et al., 2018), (Wolf-Sonkin et al., 2018), (Kirov and Cotterell, 2018), (Faruqui et al., 2016), (Faruqui et al., 2015), (Aharoni and Goldberg, 2016), (Cotterell et al., 2017). Much of this work assumes availability of partially labelled data, such as word paradigms and/or clean datasets, such as lists of 'headwords' (lemmas) from which paradigms are generated. (Fam and Lepage, 2018) identify three main approaches to learning morphological inflection: the hand-engineered rule-based approach, which requires much cost and time for construction, the supervised approach, which relies on initial labelled datasets and the neural approach, which needs more training time and even more data. However, for low-resource scenarios more attention need to be given to unsupervised knowledge-light methods, which could make strong assumptions, e.g., based on compact linguistic descriptions of the inflection systems, but for the most part rely only unlabelled data or resources that would be typically available for low-resource languages.

A terminological note: in several papers, such as (Ahlberg et al., 2015), (Silfverberg et al., 2018), the term 'paradigm' is used to describe a generalised inflection pattern, which could apply to a class of words, while the term 'inflection table' characterises an individual system of inflection for a single word. This usage differs from the traditional understanding of the notion of a paradigm as a system of word forms for a given word, see e.g., (Spencer, 2001). In this paper I adhere to the traditional terminological usage for the term 'paradigm' as a system of word forms, and use the term 'inflection tables' referring only to tables of inflections, which may be attached to a class of stems.

The problem of characterising dynamic processes in the Ukrainian lexicon has been discussed in (Klymenko et al., 2008), (Karpilovs'ka et al., 2008), where these changes are attributed to political, cultural and technical developments in the society – the active 'social dynamics', which causes the active 'linguistic dynamics': renewal and additions to nominative and communicative resources of the language and changes in linguistic norms. While the grammar or phonology remain more conservative, the lexicon is very open to such changes. There is an ongoing work to record these lexical developments for Ukrainian and other languages, however, so far there is no systematic computational linguistic framework for modelling morphological features and inflections for neologisms and new Named Entities.

3 Algorithm description

The proposed algorithm uses small set of inflection tables for inflected parts-of-speech which accepts new entities (i.e., nouns, adjectives and verbs, but not numerals or pronouns, which are closed class entities) and unannotated corpus (or a frequency list compiled from such a corpus). It attempts to split each token in the corpus into its stem and inflection using all inflections in all available inflection tables. When a split is successful, it generates a full hypothetical paradigm consistent with the split, using the identified stem and all other inflections for the given table. Then these hypotheses are checked against available word forms in the corpus: whether a sufficient number of forms can be found to confirm the hypothetical paradigm. In this approach for the paradigms to be generated reliably, the new entities need to have a sufficient morphological

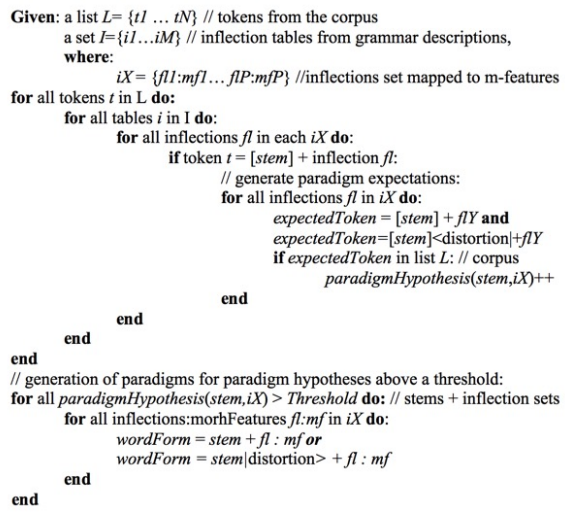

Figure 1: Algorithm description.

diversity in the corpus: it has been experimentally established that at least 3 or 4 different word forms are needed to make a reasonably accurate prediction of a paradigm and its remaining unseen word forms. For confirmed paradigms the algorithm generates all remaining word forms, their lemma (as a designated 'dictionary' word form in the paradigm, e.g., the nominative singular form for nouns) and their sets of morphological category values associated with inflections, based on the expected structure of the paradigm. Multiple splits of a token are possible, so hypothetical paradigms are ranked by the number of confirmed word forms, and the paradigm with the highest number is selected among the competing paradigms. Figure 1 shows the general overview of the algorithm. Scripts are released on GitHub repository: https://github.com/bogdanbabych/paralex4morphosyntax.

For a general case to cover less regular paradigms with stem alternations the algorithm may be complemented with a distortion model which modifies tested tokens and hypothetical word forms according to morphonological rules of the language, for Ukrainian this would cover historical alternations such as [o,e] -> [i] in 'newly closed' syllables, e.g., [kon'a] (horse.Gen.sing) -> [kin'] (horse.Nom.sing), [h, k, x] before [i] -> [z, ts, s], e.g., [ruka] (hand.Nom.sing) -> [rutsi] (hand.Dat.sing), etc.

The example in Figure 2 illustrates working of the algorithm. In this example I assume that the current token is рука (ruka = 'hand'), for which the algorithm will try to generate paradigms. In this dataset I have inflection tables for the 'hard', 'soft', 'iotated' and 'mixed' groups of the 1st declination of nouns, taken from a Ukrainian grammar textbook (Hryshchenko et al., 1997): фабрика, робітниця, надія, площа (fabryka = 'factory', robitnyts'a = 'worker', nadija = 'hope', ploshcha = 'town square'): I use only inflection sets and morphological values from these tables (the stems in the inflection tables are only for illustration).

In the first stage the algorithm tries every inflection in every table to split the current token ('ruka'). A possible split is found in the inflection table for the 1st declination of nouns, for the 'hard' and 'mixed' groups illustrated by examples 'fabryka' and 'ploshcha'. The split separates the stem *'ruk'* and the inflection *'a'*.

In the second stage, trying the split for the 'hard' group, the word form hypotheses are generated from the inflection table: *ruk+y, ruk+u, ruk+oju, ruk-o, ruk, ruk+amy, ruk+ax* and with the distortion model: [k] /_[i] -> [ts] – *ruts+i*. For the split defined by the 'mixed' group inflection table, in addition two incorrect word forms will be generated **ruk+eju, *ruk-e*, but the following three correct forms will not be generated: *ruk-y, ruk-oju, ruk-o*. Therefore, two paradigms for the split *ruk|a* will be competing with each other.

In the third stage, each of the competing paradigms will be verified against the corpus: in this example, for the 'hard' group the following four hypothesised word forms are actually found: *ruk, ruk+am, ruk+ax, ruk+y*, which, together with the 5th original form *ruk+a*, correspond to 7 morphological feature combinations, since *ruk+y* is ambiguous having three interpretations. While for the 'mixed' group only three hypothesised forms will be confirmed + initial *ruk+a* = 4, because the existing form *ruk+y* has not been predicted by the 'mixed' paradigm. As a result, the correct 'hard' paradigm will be ranked higher, with 5 confirmed word form hypotheses vs. 4 confirmed hypotheses for the wrong 'mixed' paradigm. When the corpus gets larger, more clues may differentiate such closely competing paradigms and more correct rankings may be produced.

In the fourth stage the top-ranking paradigm is confirmed and previously unseen word forms are generated, as well as possible part-of-speech code, lemma and all possible morphological feature combinations for both seen and unseen word

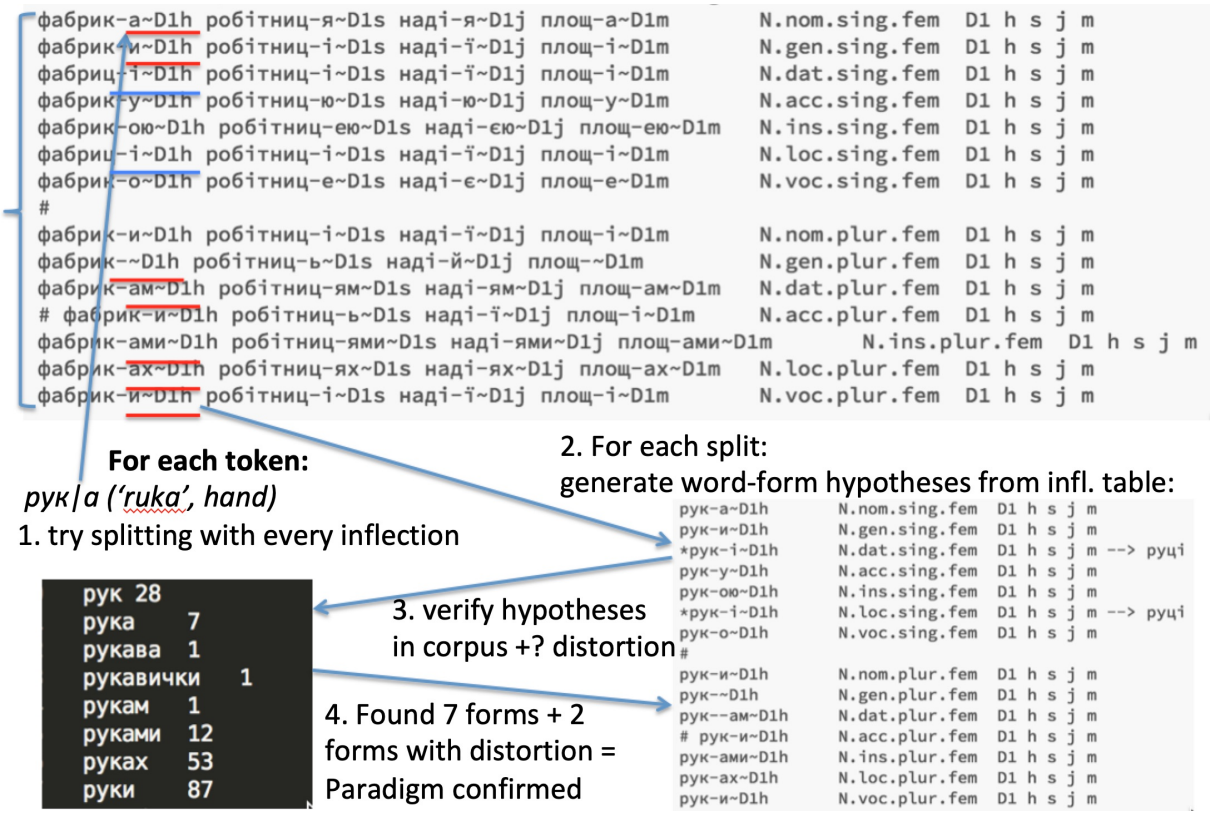

Figure 2: Illustration of the algorithm.

forms, such as values for case, number, gender, e.g., the unseen word form *ruk+u* will be generated with its morphological information *ruk-u : lem=ruka; PoS=N.acc.sing*, etc.

Note that for a single token it is not possible to clearly distinguish between a competing wrong paradigm and an alternative legitimate paradigm, which corresponds to a different reading of an ambiguous word form. For example, the word form *prykladⅼy* belongs both to *PoS=N.nom.plur; lemma=pryklad ('example')* and to *PoS=V.imper.pers2.sing; lemma=pryklasty/pryklad+u ('to attach')*. The limitation of the algorithm is that only one of these correct paradigms is confirmed for the given word form *pryklady*, depending on how many hypothesised word forms are found in corpus for each of the verbal or nominal paradigms. However, the same paradigm is confirmed via different routes, i.e., through splitting other word forms belonging to the same paradigm, e.g., in Figure 2 the paradigm for *lemma=ruka, PoS=N* will be also confirmed via splitting the corpus tokens *rukⅼam, rukⅼamy, rukⅼax*. This gives the proposed algorithm an advantage, compared to the approach described in (Ahlberg et al., 2015) for the case of overlapping paradigms with ambiguous word forms: alternative readings will be confirmed by the tokens in corpus which are unique for each of the alternative paradigms, e.g., *pryklad+ut' ('they will attach')* vs. *pryklad+om ('with an example')*. Interestingly, ambiguous word forms are not discarded: when unambiguous tokens are split for each of the overlapping paradigms, the ambiguous tokens will count in both cases to confirm both of the correct paradigms. This will not happen for competing wrong paradigms: their wrong word form predictions (such as **ruk+eju* in the example above) will simply not be found in corpus, so they will not initiate the process for the alternative paradigm.

For the purposes of this experiment I evaluate the coverage given by the algorithm without a distortion model, as such alternations are more typical for the older lexicon and often do not occur in recently borrowed items, e.g. [portu] (port.Gen.sing) - port (port.Nom.sing): [o] -> [i] alternation does not take place, as the word was borrowed after the phonological law of the 'open syllable' no longer worked in Ukrainian. However, in future the distortion models may be learnt from data or directly coded as explicit linguistic

Corpus	No of generated word forms
dict_uk	5,690,688
News	3,292,591
Wikipedia	3,765,774
Fiction	958,233
Law	1,788,288
All corpora	6,626,004

Table 2: Size of lexicon extracted from corpora.

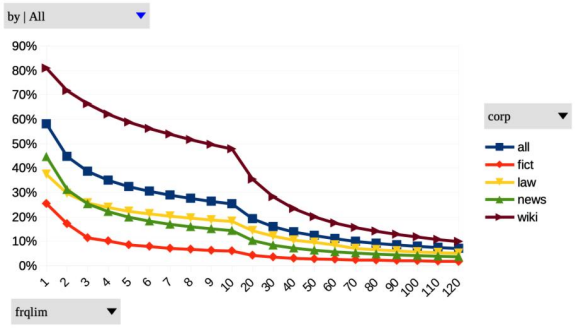

Figure 3: The baseline: Percent of non-covered types (y) in dict_uk lexicon, with filtered out lower frequencies, up to (x).

knowledge, and in this way the older paradigms and live stem alternations may also be covered.

4 Evaluating algorithm with corpus coverage

The algorithm is used for extending the Ukrainian morphological lexicon from four corpora: news, wikipedia, law and fiction, and from a combined corpus that merges these four corpora. Table 2 shows the number of word forms extracted from each of the corpora presented in Table 1.

We measure the coverage (in terms of lexical types) in four corpora and in the merged corpus, with gradually filtering out lower frequency ranges. The rationale for this evaluation method is that it is usually harder for the lexicon to cover low-frequent items, so I test this lexicons on a range of tasks of varying difficulty.

Another important aspect of evaluation would be the accuracy of the generated paradigms, e.g., the proportion of correctly generated entries, which in this paper is evaluated only indirectly, as the coverage on previously unseen corpus, as correctly generated paradigms should cover more types in the unseen corpora. Direct evaluation of accuracy will be a matter of future work, as it requires systematic sampling for different frequency ranges and more extensive manual annotation effort, which is beyond the scope of this paper.

Note that the accuracy evaluation for the algorithm would require a more complex potentially multidimensional metric, which would need to address the following aspects of accuracy: (a) correctness of the whole paradigm vs. correctness of individual forms and morphological codes, such as the case labels for animate vs. inanimate nouns that, e.g., may overlap in accusative and genitive or nominative, depending on this morphological category; (b) partial overlaps of sub-paradigms, e.g., soft, hard and mixed phonological groups, or the regular masculine vs. neuter overlap in several indirect case values in nominal and adjectival paradigms, etc.; (c) defective paradigms and potential word forms; (d) morphological variants in paradigms; (e) word forms determined by independent parameters but not by the paradigm-wide inflection class, e.g., vocative case in Ukrainian; (f) different impact of errors in paradigms, depending on syntagmatic frequency, which may be determined stylistically, grammatically or lexically (e.g., the imperative is less frequent in narrative texts, so imperative errors should be counted as less serious then errors in more common 3rd-person-singular forms.

Even though the accuracy evaluation would be important for understanding theoretical value of the proposed approach, it is less relevant for the practical scenario of updating the morphological lexicon for specialised domain, compared to the evaluation of coverage: incorrect (overgenerated) word forms in paradigms normally should not cause any additional errors compared to the baseline, as they would simply not match, the same as without the added lexicon.

As the baseline, Figure 3 shows the coverage of the existing static lexicon from the dict_uk project developed by (Rysin and Starko, 2019) (also characterised in the first row in Table 2). The horizontal axis indicates which frequency range has been filtered out. (Note the change of scale in the middle of the graph from 1 to 10 in one unit of length).

It can be seen from the figure that the Wikipedia corpus that contains many Named Entities and specialised terminology is the most problematic in terms of coverage: up to 80% of its types are not covered, which goes down only to around 50% if the frequency threshold is reduced to 10. At the same time a 'static' corpus of fiction texts is cov-

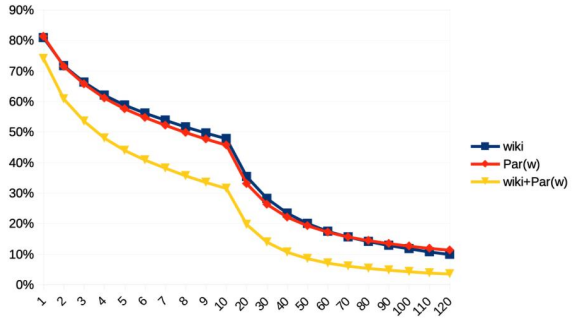

Figure 4: Percent of non-covered types (y) in Wiki corpus, with dict_uk ('wiki' line), with only proposed algorithm and paradigms generated from News corpus ('Par(w)' line), and with the two morphological lexicons combined (wiki+Par(w) line); filtered out lower frequencies, up to (x).

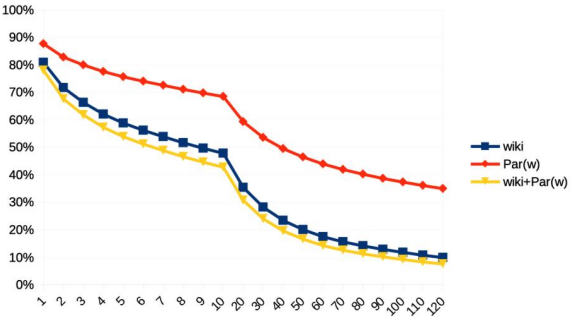

Figure 5: Percent of non-covered types (y) in Wiki corpus, with dict_uk ('wiki' line), with only proposed algorithm and paradigms generated from Law corpus ('Par(w)' line), and with the two morphological lexicons combined (wiki+Par(w) line); filtered out lower frequencies, up to (x).

ered the best by the existing morphological lexicon.

We evaluate the effect of the proposed algorithm via measuring improvements in coverage of lexical types across the frequency ranges for filtered out items. I use different corpora for the development and evaluation, so the following figures show the corpus coverage for these different combinations (lower lines indicate better results). Figure 4 and Figure 5 show coverage levels for the Wiki corpus with paradigms generated from the News and Law corpora respectively. Figure 6 and Figure 7 show coverage for the News corpus with paradigms developed from the Wiki and Law corpora. Finally, Figure 8 and Figure 9 show coverage for the Law corpus with paradigms developed from the Wiki and News corpora. In these figures the baseline graphs labelled 'wiki', 'news' and 'law' are the same as shown in Figure 3.

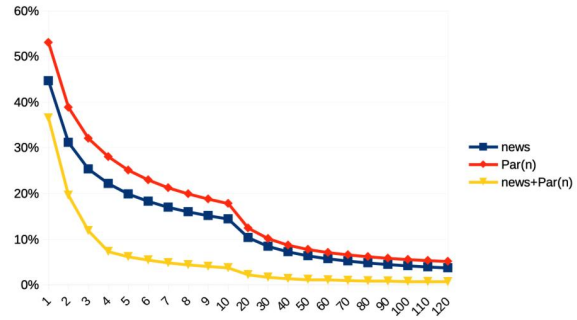

Figure 6: Percent of non-covered types (y) in News corpus, with dict_uk ('news' line), with only proposed algorithm and paradigms generated from Wiki corpus ('Par(n)' line), and with the two morphological lexicons combined (news+Par(n) line); filtered out lower frequencies, up to (x).

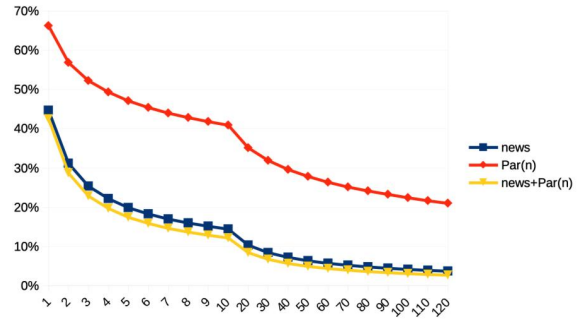

Figure 7: Percent of non-covered types (y) in News corpus, with dict_uk ('news' line), with only proposed algorithm and paradigms generated from Law corpus ('Par(n)' line), and with the two morphological lexicons combined (news+Par(n) line); filtered out lower frequencies, up to (x).

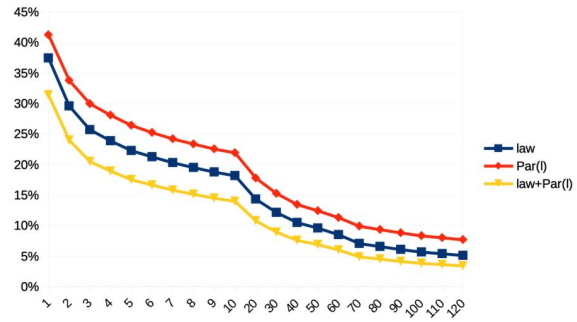

Figure 8: Percent of non-covered types (y) in Law corpus, with dict_uk ('law' line), with only proposed algorithm and paradigms generated from Wiki corpus ('Par(l)' line), and with the two morphological lexicons combined (law+Par(l) line); filtered out lower frequencies, up to (x).

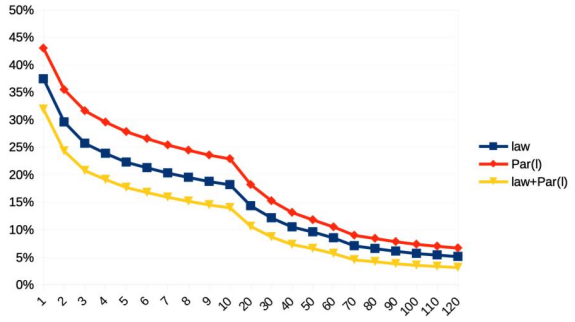

Figure 9: Percent of non-covered types (y) in Law corpus, with dict_uk ('law' line), with only proposed algorithm and paradigms generated from News corpus ('Par(l)' line), and with the two morphological lexicons combined (law+Par(l) line); filtered out lower frequencies, up to (x).

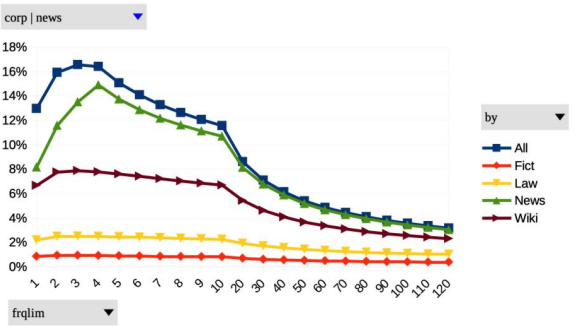

Figure 10: Improvement rates (y), of different corpora with the paradigms generated from New corpus; filtered out lower frequencies, up to (x).

It can be seen from the figures that for the proposed algorithm the morphological and lexical diversity of the corpus are essential: the Law corpus has very little effect on the coverage of both News and Wikipedia corpora, while the News and Wikipedia consistently improve the coverage of all the corpora on which they are evaluated. This may be due to the small type/token ratio (i.e., small lexical diversity) of the Law corpus.

Finally, Figures 10 and 11 summarise improvement rates (i.e., the difference between the baseline and the proposed approach) for all the corpora using the News and Wiki corpora for generating paradigms.

It can be seen from these figures that the coverage of Fiction and Law corpora is harder to improve, while News and Wiki corpora are most complementary, improving each other well. Also an interesting effect can be observed when a corpus is used to improve itself: the improvement rate peaks at the value of filtered frequencies up to 4,

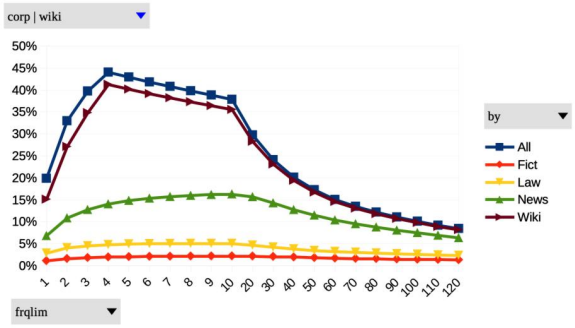

Figure 11: Improvement rates (y), of different corpora with the paradigms generated from Wiki corpus; filtered out lower frequencies, up to (x).

which may be interpreted as improvement in reliability of paradigm prediction for more frequent items, and as an indication of a possible threshold for the minimal number of slots for predicting Ukrainian paradigms.

The results indicate that the proposed approach gives consistent improvements in coverage with paradigms generated from a lexically diverse corpora with sufficient number of neologisms and new proper names. The highest improvement rates across different corpora has been achieved for paradigms generated from the Wiki corpus used to test the News corpus – 16.3% for lexical items with frequencies 10 and higher.

5 Discussion

The algorithm proposed in this paper uses the fundamental idea that "the existence of a hypothetical lemma can be guessed if several different words found in the corpus are best interpreted as morphological variants of this lemma" (Clément et al., 2004). This idea has been developed for automated induction of morphological lexica for different languages and implemented in practical applications such as spell checking (e.g., ispell) and information retrieval systems: (Krovetz, 1993), (Grefenstette et al., 2002), (Segalovich, 2003), (Clément et al., 2004), (Oliver and Tadić, 2004), (Sagot, 2005); recent work in this area uses more accurate machine learning approaches: (Šnajder, 2013), (Ljubešić et al., 2015).

The approach proposed in this paper develops these ideas further by explicitly focussing on the following conceptual points:

(1) The extracted units are paradigms, and not lemmas or mappings from inflected word forms to lemmas or paradigms. The advantage of such

approach is that inflected word forms in the corpus provide only indirect, latent justification for the existence of paradigms, so there is a separation between a form which initiates generation of hypotheses and those (possibly ambiguous) word forms that are used as evidence: they can independently justify the existence of several different paradigms, which avoids the artificial pressure to choose a single top-ranked paradigm and lemma for each inflected form, as it is the case in (Oliver and Tadić, 2004) or (Ahlberg et al., 2014). This latent paradigm induction is also more robust against potential noise in the corpus, since misspellings would not normally collect enough inflected forms.

(2) The proposed approach focusses only on the regular dynamic component of the lexicon, which enables a clean separation between the core method of the paradigm induction and the extensions or other methods needed to address historical or irregular features, such as stem alternations or suppletive forms, for which separate distortion models can be developed or learnt from corpora. Also the inflection tables for generating paradigms hypotheses are derived from comprehensive grammatical descriptions rather than from potentially noisy data. This reflects a typical scenario of morphological lexicon development for many under-resourced languages, which still have a smaller dictionary that covers most frequent items and comprehensive inflection tables in traditional grammars, but where it is hard to recruit language specialists on a recurrent basis to keep up with constant lexical developments in different subject domains of the language for which applications need to be developed or updated.

(3) Evaluation in the proposed approach is part of the development workflow: it focusses on the dynamics of corpus coverage with generated word forms for different maximum frequency thresholds. Such comprehensive automated evaluation indicates on the large scale where maximal improvement in coverage can be expected, so which frequencies can be used as cut-off points to filter out noisier and less reliable paradigms.

Most lexical items covered with the proposed paradigm generation algorithm are single-word Named Entities – names of organisations, geographical places or people, as well as technical terms, e.g.: мінохоронздоров'я ('The Ministry of Health') інтербізнесконсалтинг ('Internet business consulting'), кременчукм'ясо ('The Meat of Kremenchuk' company), кривбасводопостачання ('Kryvbas Water Supply'), броварітепловодоенергія ('Brovary Heating, Water and Energy' company), могадішо ('Mogadishu') озоноруйнуючих . ('ozone-destroying').

However, the list also contains interesting political lexicon, such as йолка ('Christmas tree': the distorted ukrainized spelling of the Russian word, which became a symbol of the people's resistance to political violence during the Ukrainian revolution of dignity in 2013-2014) and профферсор (again, a distorted spelling of the word 'professor', which was used for mocking the fugitive pro-Russian president, who held this title, but allegedly misspelt it in an official document).

The appearance of this politically charged lexicon is in line with Karpilovs'ka et al.'s (2008) suggestion that lexical changes are driven by the social dynamics, especially at the times of major political developments. However, it can be also seen that this political lexicon is still much less frequent and less changeable compared to Named Entities, which dominate the new lexicon.

6 Conclusions and future work

The proposed algorithm complements static linguistic resources and increases corpus coverage for new entities, such as neologisms and proper names. The highest improvements are achieved for the corpus types that typically have many neologisms, specialised terminological lexicon and Named Entities: the Wikipedia and News. These corpora are not well covered by existing morphological resources. The advantage of the proposed approach is that it uses unlabelled corpora and small inflection tables for unsupervised induction of paradigms. However, its limitation is that in this stage it doesn't predict irregular paradigms.

Future work will involve the development of distortion models to cover less regular cases and a systematic evaluation of the accuracy of paradigm prediction for different frequency ranges: while for more frequent items such prediction is highly reliable, there is a need to experimentally establish frequency and coverage thresholds for different error rates on this task for less frequent items. Another area for future research is the use of contextual and syntactic features to verify predicted morphological properties.

References

Roee Aharoni and Yoav Goldberg. 2016. Morphological inflection generation with hard monotonic attention. *arXiv preprint arXiv:1611.01487*.

Malin Ahlberg, Markus Forsberg, and Mans Hulden. 2014. Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 569–578.

Malin Ahlberg, Markus Forsberg, and Mans Hulden. 2015. Paradigm classification in supervised learning of morphology. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1024–1029.

Bogdan Babych and Serge Sharoff. 2016. Ukrainian part-of-speech tagger for hybrid mt: Rapid induction of morphological disambiguation resources from a closely related language. In *Fifth Workshop on Hybrid Approaches to Translation (HyTra)*. EAMT, Riga: June 1st 2016.

Lionel Clément, Bernard Lang, and Benoît Sagot. 2004. Morphology based automatic acquisition of large-coverage lexica. In *LREC 04*, pages 1841–1844.

Ryan Cotterell, Ekaterina Vylomova, Huda Khayrallah, Christo Kirov, and David Yarowsky. 2017. Paradigm completion for derivational morphology. *arXiv preprint arXiv:1708.09151*.

Vsevolod Dyomkin, Dmytro Chaplinskyi, Anatoliy Stegnii, Oleksandr Marikovskyi, Viacheslav Tykhonov, Oles Petriv, Serhii Shekhovtsov, Mykhailo Chalyi, Tetiana Kodliuk, Mykyta Pavliuchenko, Oksana Kunikevych, and Khrystyna Skopyk. 2019. lang-uk. *GitHub repository*. http://lang.org.ua/en/corpora/#anchor4.

Rashel Fam and Yves Lepage. 2018. Ips-waseda system at conll-sigmorphon 2018 shared task on morphological inflection. *Proceedings of the CoNLL SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 33–42.

Manaal Faruqui, Ryan McDonald, and Radu Soricut. 2016. Morpho-syntactic lexicon generation using graph-based semi-supervised learning. *Transactions of the Association for Computational Linguistics*, 4:1–16.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2015. Morphological inflection generation using character sequence to sequence learning. *arXiv preprint arXiv:1512.06110*.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Gregory Grefenstette, Yan Qu, and David Evans. 2002. Expanding lexicons by inducing paradigms and validating attested forms. In *LREC 2002*.

Tetiana O. Gryaznukhina, editor. 1999. *Syntactic analysis of scientific texts on computers. / Sintaksicheskiy analiz nauchnogo teksta na EVM*. Naukova Dumka, Kyiv, Ukraine.

Arnold P. Hryshchenko, Matsko Liubov I., Pliushch Mariya Ja., Totska Nina I., and Uzdyhan Ivanna M. 1997. *Modern Ukrainian Literary Langauge / Suchasna Ukrajins'ka Literaturna Mova*. Vyshcha Shkola, Kyiv.

Yevheniya A. Karpilovs'ka, Larysa P. Kysliuk, Nina F. Klymenko, Valentyna I. Kryts'ka, Puzdyr'eva Tetiana K., and Yuliya V. Romaniuk. 2008. *Active resources of modern Ukranian nomination. Ideographic dictionary of the new lexicon / Aktyvni resursy suchasnoji Ukrajins'koji nominatsiji: Ideohrafichnyj slovnyk novoji leksyky*. KMM, Kyiv.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Nina F. Klymenko, Yevheniya A. Karpilovs'ka, and Larysa P. Kysliuk. 2008. *Dynamic processes in the modern Ukrainian lexicon / Dynamichni protsesy v suchasnomu ukrajins'komu leksykoni*. Vydavnychyi Dim Dmytra Burago.

Kimmo Matti Koskenniemi et al. 2018. Guessing lexicon entries using finite-state methods. In *Proceedings of the Fourth International Workshop on Computatinal Linguistics for Uralic Languages*. Association for Computational Linguistics.

Natalia Kotsyba, Andriy Mykulyaky, and Igor Shevchenko. 2009. Ugtag: morphological analyzer and tagger for the ukrainian language. In *Proceedings of the international conference Practical Applications in Language and Computers (PALC 2009)*.

Natalia Kotsyba, Igor Shevchenko, Ivan Derzhanski, and Andriy Mykulyak. 2010. Multext-east morphosyntactic specifications, version 4. 3.11 url: http://nl.ijs.si/ME/V4/msd/html/msd-uk.html.

Robert Krovetz. 1993. Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 191–202. ACM.

Nikola Ljubešić, Miquel Espla-Gomis, Filip Klubička, and Nives Mikelić Preradović. 2015. Predicting inflectional paradigms and lemmata of unknown words for semi-automatic expansion of morphological lexicons. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 379–387.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *LREC*.

Antoni Oliver and Marko Tadić. 2004. Enlarging the croatian morphological lexicon by automatic lexical acquisition from raw corpora. In *Fourth International Conference on Language Resources and Evaluation LREC2004*. ELRA.

Valentyna S. Perebyjnis, Nataliya P. Darchuk, and Tetiana O. Gryaznukhina. 1989. *Morphological analysis of scientific texts on computers / Morfologicheskiy analiz nauchnogo teksta na EVM*. Naukova Dumka, Kyiv, Ukraine.

Ian Press and Stefan Pugh. 2015. *Ukrainian: A comprehensive grammar*. Routledge.

Aarne Ranta. 2011. *Grammatical framework: Programming with multilingual grammars*, volume 173. CSLI Publications, Center for the Study of Language and Information Stanford.

Andriy Rysin and Vasyl Starko. 2019. Project to generate pos tag dictionary for ukrainian language. *GitHub repository*. https://github.com/brown-uk/dict_uk.

Benoît Sagot. 2005. Automatic acquisition of a slovak lexicon from a raw corpus. In *International Conference on Text, Speech and Dialogue*, pages 156–163. Springer.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA*, pages 273–280. Citeseer.

Miikka Silfverberg, Ling Liu, and Mans Hulden. 2018. A computational model for the linguistic notion of morphological paradigm. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1615–1626.

Jan Šnajder. 2013. Models for predicting the inflectional paradigm of croatian words. *Slovenščina*, 20:1–34.

Andrew Spencer. 2001. The paradigm-based model of morphosyntax. *Transactions of the Philological Society*, 99(2):279–314.

Ludwig Wittgenstein. 2009. *Philosophical investigations*. Wiley-Blackwell.

Lawrence Wolf-Sonkin, Jason Naradowsky, Sebastian J Mielke, and Ryan Cotterell. 2018. A structured variational autoencoder for contextual morphological inflection. *arXiv preprint arXiv:1806.03746*.