1

Choosing the correct paradigm for unknown words in rule-based machine translation systems

V. M. Sánchez-Cartagena, M. Esplà-Gomis, F. Sánchez-Martínez, J. A. Pérez-Ortiz

Abstract

Previous work on an interactive system aimed at helping non-expert users to enlarge the monolingual dictionaries of rule-based machine translation (MT) systems worked by discarding those inflection paradigms that cannot generate a set of inflected word forms validated by the user. This method, however, cannot deal with the common case where a set of different paradigms generate exactly the same set of inflected word forms, although with different inflection information attached. In this paper, we propose the use of an n-gram-based model of lexical categories and inflection information to select a single paradigm in cases where more than one paradigm generates the same set of word forms. Results obtained with a Spanish monolingual dictionary show that the correct paradigm is chosen for around 75% of the unknown words, thus making the resulting system (available under an open-source license) of valuable help to enlarge the monolingual dictionaries used in MT involving non-expert users without technical linguistic knowledge.

Choosing the correct paradigm for unknown words in rule-based machine translation systems.
Víctor M. Sánchez-Cartagena,
Miquel Esplà-Gomis,
Felipe Sánchez-Martínez,
Juan Antonio Pérez-Ortiz,
Departament de Llenguatges i Sistemes Informàtics,
Universitat d'Alacant, Spain.

1.1 Introduction

Rule-based machine translation (MT) systems heavily depend on explicit linguistic data such as monolingual dictionaries, bilingual dictionaries, grammars, etc. (Hutchins and Somers, 1992). Although the acquisition of these data has usually required the intervention of linguists, development costs could be significantly reduced by involving a broader group of non-expert users in the enrichment of these MT systems. This group may include, for instance, people using an online translation service who want to improve it by adding an unknown word to the underlying MT system dictionaries, or collaborators recruited through crowdsourcing (Wang et al., 2010) platforms.

In previous works (Esplà-Gomis et al., 2011, Sánchez-Cartagena et al., 2012) we proposed a novel method for enlarging the two monolingual dictionaries and the bilingual dictionary of shallow-transfer rule-based MT systems with the collaboration of non-expert users. In the case of a monolingual dictionary, adding a new entry implies determining the stem of the new word and a suitable inflection *paradigm* among those defined by the MT system for the corresponding language. Paradigms are commonly introduced to group regularities in inflection which are common to a set of words; for instance, the paradigm assigned to many common English verbs, indicates that by adding *-ing* to the stem, the gerund is obtained; by adding *-ed*, the past is obtained; and so on. In our approach, the most appropriate stem/paradigm combination is chosen by means of a sequence of simple yes/no questions whose answer only requires *speaker-level* understanding of the language. Basically, users are asked to validate whether the forms resulting from tentatively assigning different candidate paradigms to the new word to be inserted are correct inflected forms of it. The experiments we performed showed (Esplà-Gomis et al., 2011) that the average number of queries posed to the users for a Spanish monolingual dictionary was 5.2, which is reasonably small considering the 56.4 initial compatible paradigms on average.

The whole procedure for adding an unknown word and its translation to all the MT system dictionaries could consequently consist of requesting users a source-language word and its corresponding translation into target language (for instance, *cars* and *coches*, for an English–Spanish MT system). Then, our method could be independently applied to the source-language word and its target-language translation to obtain their inflection paradigms and insert all this information into the monolingual dictionaries. Finally, the corresponding link between both words could be inserted in the bilingual dictionary in a straightforward

manner without any additional user interaction. Moreover, we have shown (Sánchez-Cartagena et al., 2012) that when the source-language word has been already inserted, the system is able to more accurately predict the right target-language paradigm by exploiting the correlations between paradigms in both languages, thus reducing significantly the number of queries posed to the user. Note that, although when the source language and the target language are not closely related the correlation between paradigms is not very strong, previous experiments performed with the English–Spanish language pair (Sánchez-Cartagena et al., 2012) have shown that the source-language part of speech is still useful to reduce the number of queries posed when inserting the target-language word.

Our proposal provided a complete framework for dictionary enlargement, but it still lacked a critical component to discriminate between paradigms providing the same set of inflected word forms. It turns out that by only asking users whether a set of word forms are correct forms of the word to be inserted, our system frequently ends up with more than one feasible stem/paradigm solution and, since all of them generate the same set of inflected word forms, no additional query can be posed to the user in order to discriminate between them. For example, in the case of Spanish, the inflected word forms for many adjectives such as *alto* (*tall* in English) are *alt/o* (masculine, singular), *alt/a* (feminine, singular), *alt/os* (masculine, plural), *alt/as* (feminine, plural);[1] therefore, dictionaries contain a paradigm for adjectives with suffixes {*-o,-a,-os,-as*} to which the stem *alt-*, among others, will be assigned. Additionally, the inflected word forms for many nouns such as *gato* (*cat* in English) are *gat/o* (masculine, singular), *gat/a* (feminine, singular), *gat/os* (masculine, plural), *gat/as* (feminine, plural); consequently, dictionaries contain a paradigm for nouns with suffixes {*-o,-a,-os,-as*} to which stems such as *gat-* will be assigned. As can be seen, in the case of adding an unknown word such as the noun *perro* (*dog* in English), which inflects as *gato*, no yes/no question may be presented to the user to discriminate between the two paradigms (which are equivalent for the interactive method) given the stem *perr-*.

Note that the problem shows similarities to that of part-of-speech tagging (Manning and Schütze, 1999) and it can be addressed through similar approaches, but in our case we also need to disambiguate between equivalent paradigms involving the same lexical category. For instance, the Spanish inflected forms for nouns such as *abeja* (*bee* in

[1] In this paper, we use the slash character to separate the stem of a word from the suffix of one of its possible inflections.

English) are *abeja/*ϵ[2] (feminine, singular), *abeja/s* (feminine, plural); and the inflected word forms for the noun *abismo* (*abyss* in English) comprise *abismo/*ϵ (masculine, singular), *abismo/s* (masculine, plural). Therefore, these two words are assigned to equivalent paradigms, both for the noun lexical category with suffixes {ϵ,-*s*}, but with different inflection information (gender). A new noun such as *taza* (*cup* in English) would in principle fit into both paradigms.

Although in this paper we give a fully automatic solution to the multiple step/paradigm issue, an interactive approach could also be followed. Users may be asked to validate some sentences in which the word to be classified would contain the inflecion information from each paradigm. For instance, one possible strategy for eliciting the gender of *taza* would be to ask the user to validate the sentences *el taza* and *la taza*, being *el* a masculine determiner and *la* a feminine determiner.

Our automatic solution is obtained by introducing an n-gram-based model of lexical categories and inflection information which is used to automatically choose the right stem/paradigm combination: nouns belonging to the same paradigm as *abeja* will be usually preceded by a feminine determiner in a corpus, whereas nouns to be assigned to the same paradigm as *abismo* will be frequently preceded by a masculine determiner.

The model is trained with a monolingual corpus where every word is replaced by its morphological analysis comprising lexical category and inflection information. The Java code for the resulting system is available under the free/open-source GNU General Public License[3] and may be downloaded from `https://apertium.svn.sourceforge.net/svnroot/apertium/branches/dictionary-enlargement`.

In the experiments we have used the free/open-source rule-based MT system Apertium (Forcada et al., 2011), which is being currently used to build MT systems for a large variety of language pairs. In the case of the Spanish monolingual dictionary used in the Spanish–Catalan Apertium MT system, 81.1% of the words would be assigned by the original method to more than one equivalent paradigm; as a result, giving a solution to the multiple paradigm issue is critical.

The rest of the paper is organised as follows. Section 1.2 discusses other works related to our proposal. Section 1.3 introduces some concepts about monolingual dictionaries which will be used in the rest of the paper. An overview of the previous method for dictionary enlargement is presented in section 1.4, followed by the description of

[2]Symbol ϵ denotes the empty string.

[3]`http://www.gnu.org/licenses/gpl.html`

our new improvement for discriminating between paradigms generating the same inflected forms in section 1.5. Section 1.6 discusses our experimental setting. Then, the results obtained are presented and discussed in section 1.7. Finally, some concluding remarks are presented in section 1.8.

1.2 Related work

Two of the more prominent works related to the elicitation of knowledge for building or improving MT systems are those by Font-Llitjós (2007) and McShane et al. (2002). The former proposes a strategy for improving both transfer rules and dictionaries by analysing the postediting process performed by a non-expert user through a special interface. McShane et al. (2002) design a complex framework to elicit linguistic knowledge from informants who are not trained linguists and use this information to build MT systems which translate into English; their system provides users with a lot of information about different linguistic phenomena to ease the elicitation task.

Unlike the Avenue formalism used in the work by Font-Llitjós (2007), the MT system we are using is a *pure* transfer-based one in the sense that a single translation is generated and no language model is used to score a set of possible candidate translations; therefore, we are interested in a single correct solution and assume that an incorrect paradigm cannot be assigned to a new word. Unlike the works by McShane et al. (2002) or Bartusková and Sedlácek (2002), we want to relieve users of acquiring linguistic skills.

1.3 Monolingual dictionaries in rule-based MT systems

As already pointed out, monolingual dictionaries have two types of data: *paradigms*, that group regularities in inflection, and *word entries*, represented by a stem and a paradigm. The *stem* is the part of a word that is common to all its inflected variants. Paradigms make easier the management of dictionaries in two ways: by reducing the quantity of information that needs to be stored, and by simplifying revision and validation thanks to the explicit encoding of regularities in the dictionary. Once the most frequent paradigms in a dictionary are defined, entering a new word is generally limited to writing the stem and choosing an inflection paradigm. In this work we assume that all the paradigms for the words in the language are already included in the dictionary.

Let $P = \{p_i\}$ be the set of paradigms in a monolingual dictionary. Each paradigm p_i defines a set F_i of pairs (f_{ij}, m_{ij}), where f_{ij} is a

suffix[4] which is appended to stems to build new *inflected word forms* (IWFs), and m_{ij} is the corresponding morphological information.

Given a *stem/paradigm* pair c composed of a stem t and a paradigm p_i, the *expansion* $I(t, p_i)$ is the set of possible IWFs resulting from appending each of the suffixes in F_i to t. For instance, an English dictionary may contain the stem *want-* assigned to a paradigm with suffixes[5] $F_i = \{\epsilon$,*-s*, *-ed*, *-ing*$\}$ (ϵ denotes the empty string); the expansion $I(\text{want}, p_i)$ consists of the set of IWFs *want*, *wants*, *wanted* and *wanting*. We also define a *candidate stem* t as an element of $\Pr(w)$, the set of possible prefixes of a particular IWF w.

1.4 Original method

As our new proposal is a refinement over our previous method (Esplà-Gomis et al., 2011) for adding new entries to the monolingual dictionaries of an MT system, a brief description of it follows before presenting the main contribution of this paper in section 1.5.

Given a new IWF w to be added to a monolingual dictionary, our objective is to find both the candidate stem $t \in \Pr(w)$ and the paradigm p_i which expand to the largest possible set of IWFs which are correct forms of w. To that end, our method performs these three tasks: paradigm detection, paradigm scoring, and user interaction.

Paradigm detection. To detect the set of paradigms which may produce the IWF w and their corresponding stems we use a *generalised suffix tree* (McCreight, 1976) containing all the possible suffixes included in the paradigms in P. A list L is built containing all the candidate stem/paradigm pairs compatible with the IWF to be added (candidate paradigms, CPs). We will denote each of these candidates as c_n.

The following example illustrates this stage of our method. Consider a simple dictionary with only four paradigms: p_1, with F_1=$\{f_{11}$=ϵ, f_{12}=*-s*$\}$; p_2, with F_2=$\{f_{21}$=*-y*, f_{22}=*-ies*$\}$; p_3, with F_3=$\{f_{31}$=*-y*, f_{32}=*-ies*, f_{33}=*-ied*, f_{34}=*-ying*$\}$; and p_4, with F_4=$\{f_{41}$=*-a*, f_{42}=*-um*$\}$. Let's assume that a user wants to add the new IWF w=*policies* (actually, the noun *policy*) to the dictionary. The candidate stem/paradigm pairs which will be obtained after this stage are: c_1=*policies*/p_1; c_2=*policie*/p_1; c_3=*polic*/p_2; and c_4=*polic*/p_3.

[4] Although our approach focuses on languages generating word forms by adding suffixes to the stems of words (for example, Romance languages), it could be easily adapted to inflectional languages based on different ways of adding morphemes as long as this kind of inflection is encoded in paradigms; note that a data structure different from a suffix tree (see section 1.4) may be needed.

[5] We omit the morphological information contained in F_i and show only the suffixes.

Paradigm scoring. Once L is obtained, a *confidence score* is computed for each CP $c_n \in L$ using a large monolingual corpus C. Candidates producing a set of IWFs which occur more frequently in the corpus get higher scores.

Following our example, the IWFs for the different candidates would be: $I(c_1)$={*policies*, *policiess*}; $I(c_2)$={*policie*, *policies*}; $I(c_3)$={*policy*, *policies*}; and $I(c_4)$={*policy*, *policies*, *policied*, *policying*}. Using a large English corpus, IWFs *policies* and *policy* will be easily found, and the rest of them (*policie*, *policiess*, *policied* and *policying*) probably will not. Therefore, c_3 would obtain the highest score.

User interaction. Finally, the best candidate is chosen from L by querying the user about a reduced set of the IWFs for some of the CPs $c_n \in L$. In this way, when an IWF w' is accepted by the user, all $c_n \in L$ for which $w' \notin I(c_n)$ are removed from L; otherwise, all $c_n \in L$ for which $w' \in I(c_n)$ are removed from L.

In order to try to maximize the number of IWFs discarded in each query and, consequently, minimize the amount of yes/no questions, our system firstly sorts L in descending order using the confidence score previously computed. Then, users are asked to confirm whether the IWF from the first CP in L which exists in the minimum number of other CPs in L is a correct form of w. This process is repeated until only one candidate or a set of equivalent paradigms remain in L.

1.5 Improvement to the method

The original method presented in the previous section has an important limitation: the system frequently ends up with more than one stem/paradigm proposal. All these final candidates generate the same set of inflected word forms, although with some variation in the lexical category or in the inflection information, and no additional query can be posed to the user in order to discriminate between them (see the introduction for some examples in Spanish).

As an empirical evidence of the importance of that limitation, we found that when trying to find the most appropriate paradigm for a representative set[6] of the words already inserted in the Spanish monolingual dictionary of the Apertium Spanish–Catalan[7] language pair, 81.1% of the entries would be assigned more than one stem/paradigm pair after users answered correctly all the queries posed by the original system described in section 1.4.

[6] See section 1.6 for details about how this set was obtained.

[7] Revision 33900 in the Apertium SVN trunk `https://apertium.svn.sourceforge.net/svnroot/apertium/trunk/apertium-es-ca`.

TABLE 1 Top 6 paradigm classes for the Spanish monolingual dictionary. Examples of inflections for the different paradigms contained in each class are given together with the total number of paradigms (# P), and the number of words (# W) in the dictionary assignable to the class. The last two columns show results described in section 1.7. Confidence intervals were estimated with 95% statistical significance with a *t-test*.

Word examples	# P	# W	Success (%)	Baseline (%)
atletismo, Suecia, adiós, afueras, ...	32	11 513	56.3 ± 1.2	40.5 ± 1.2
accionista, abeja, abismo, clarisa, abundante	5	11 507	82.9 ± 0.8	46.6 ± 1.0
abogado, cuánto, absoluto, mío, otro, todo	6	3 281	72.3 ± 1.7	78.2 ± 1.6
abdominal, abril, accesibilidad, albañil	4	2 256	87.8 ± 1.5	37.3 ± 2.2
acción, aluvión, marrón, peatón	4	2 014	96.6 ± 0.9	87.7 ± 1.6
abrumadora, señora	2	571	73.9 ± 4.0	53.1 ± 5.0

Each of the different sets of equivalent CPs which can be assigned to one or more entries in the monolingual dictionary constitute a *paradigm class*. Table 1 shows, for each of the 6 paradigm classes with most words in the Spanish dictionary, the number of entries which would be assigned to them after the original method, the number of different CPs in the final list, and an example word for every CP.

Our hypothesis is that a probabilistic model of lexical categories and inflection information could prove very useful to find the correct paradigm in the paradigm class. For instance, as already commented in the introduction, nouns belonging to the same paradigm as *abeja* will be usually preceded by a feminine determiner in a corpus, whereas nouns to be assigned to the same paradigm as *abismo* will be frequently preceded by a masculine determiner.

Consequently, we propose to train an n-gram model (Manning and Schütze, 1999) upon a monolingual corpus where every word has been replaced by its morphological analysis. In the case of the Apertium platform used in our experiments, the monolingual dictionary and the part-of-speech tagger are used to convert each inflected word form in the monolingual corpus (for instance, *abismos*), into a *lexical form* consisting of lemma, lexical category, and inflection information (*abismo,*

noun, masculine, plural); lemmas will be discarded for the purpose of our work.

The n-gram model is then used as showed in algorithm 1: for one[8] of the paradigms p_i in the paradigm class the list of all possible inflected word forms $\{w_j\}$ for the new word is obtained (function *InflectedWordForms*). Each of the word forms w_j is then sought in a monolingual corpus (function *FindSentencesContaining*), and every sentence containing w_j is morphologically analysed to obtain the lexical category and inflection information of all its words (function *ObtainLexicalForms*), except for w_j; for the occurrence of w_j in the sentence, all the possible analysis according to the different paradigms p_i in the paradigm class are tested one after the other and the perplexity per word (actually, perplexity per *token*; function PPW) of the sentence (Manning and Schütze, 1999) is computed according to the n-gram model of lexical inflection information. The paradigm containing the inflection information which makes the sentence obtain the smallest total perplexity per token is considered the winner. The process is repeated for every sentence in the corpus containing one of the forms w_j and the paradigm which is found winner more frequently is selected by the algorithm as the correct one. Note that an ordered list of all the paradigms in the paradigm class could also be obtained by following this procedure. A sorted list could be useful in scenarios where a user is requested to validate the associated paradigm before finally adding the new entry to the dictionary; in this case, if the first candidate is not valid, then the user will move to the second one and so on; ideally, very few paradigms would need to be tested before getting to the correct one.

1.6 Experimental settings

Since the addition by non-expert users of new entries to monolingual dictionaries has already been evaluated (Esplà-Gomis et al., 2011), our experimental set-up is focused on studying the impact of our lexical model in the selection of the correct paradigm when more than one stem/paradigm candidate exist after querying the user. The evaluation can be carried out automatically by focusing on the entries already included in the dictionary which would obtain more than one CP with the original method described in section 1.4. Since those entries already have the correct paradigm assigned, it is not necessary to pose the yes/no questions to users in order to have them labelled.

[8] Note that since all the paradigms in the paradigm class are equivalent in the sense that they generate exactly the same set of IWFs, any of them could be chosen here.

Algorithm 1 Steps carried out to choose the paradigm whose part of speech and inflection information best fit the new word *nw*. The function *Init* initialises to zero the amount of sentences in which each paradigm from *paradigm_class* is the best one. Note that in function *ObtainLexicalForms* the occurrence of w_j is initially marked as an unknown word, since it does not appear in the dictionary.

```
function BestParadigm( nw, list paradigm_class, corpus,ngram_model)
    list iwfs ← InflectedWordForms(nw, paradigm_class)
    map winner_paradigms ← ∅
    Init(winner_paradigms, paradigm_class)
    for all w_j ∈ iwfs do
        list occurrences ← FindSentencesContaining(w_j, corpus)
        for all occurence ∈ occurrences do
            lex_occurrence ← ObtainLexicalForms(occurrence)
            perplexity_per_word ← ∞
            best_paradigm ← null
            for all p_i ∈ paradigm_class do
                lexical_word_form ← LexForm(p_i, w_j)
                lex_occurence_replaced ← lex_occurence.Replace(w_j, lexical_word_form)
                sample_perplexity ← PPW(lex_occurence_replaced, ngram_model)
                if sample_perplexity <= perplexity_per_word then
                    perplexity_per_word ← example_perplexity
                    best_paradigm ← p_i
                end if
            end for
            winner_paradigms[best_paradigm] =
      winner_paradigms[best_paradigm] + 1
        end for
    end for
    return arg max_p winning_paradigms[p]
end function
```

We have used the Apertium Spanish–Catalan[9] language pair, and a combination of sentences from a Spanish Wikipedia dump[10] and the Spanish version of OpenSubtitles corpus (Tiedemann, 2009) as the monolingual corpus to train the n-gram model and to search for sentences containing the inflected word forms w_j in the paradigm class (see section 1.5). The n-gram model used in the experiments is a trigram model trained with the open-source toolkit IRSTLM (Federico et al., 2008) using Witten-Bell smoothing and without pruning singleton n-grams.

Our test set contains all the word entries assigned to paradigms corresponding to open part-of-speech categories which have at least two dictionary entries assigned to them. For each word in the test set, its corresponding paradigm class is obtained by checking all the possible pairs stem-paradigm generating the same IWF set than the correct stem/paradigm pair. Our new approach is then used to select one of the paradigms which is, after that, compared to the correct one according to the dictionary. As a baseline, a simple model which selects the paradigm in the class with the largest number of entries assigned to it in the monolingual dictionary is also considered. It is worth noting that the comparison is not totally fair, since the baseline uses knowledge about the number of words assigned to each paradigm in the dictionary, which is not available for our approach.

1.7 Results

Table 1 shows the results (two last columns) for the 6 most frequent paradigm classes among the 26 different paradigm classes which were found in the Spanish dictionary. These classes include 97.0% of the entries which can be assigned more than one candidate paradigm by the original method. Paradigm classes contain between two and six paradigms, except for one of them, which comprises 32 paradigms; this large class corresponds to paradigms containing only the suffix ϵ (which is assigned to words with one single inflected form, such as proper nouns). The results obtained by our approach clearly overcome the results obtained by the baseline, except for the third class. It is worth noting that our approach can only deal with words for which any occurrence of their inflected word forms appear in the corpus. Therefore, success rate was computed only for these words both for the baseline and for our approach.

[9] Revision 33900 in the Apertium SVN trunk `https://apertium.svn.sourceforge.net/svnroot/apertium/trunk/apertium-es-ca`.

[10] `http://dumps.wikimedia.org/eswiki/20110114/eswiki-20111208-pages-articles.xml.bz2`.

With regard to the overall results involving all the 32 104 entries assigned to the 26 different paradigm classes and fulfilling the conditions enumerated in section 1.6, the average success rate was $75.7\% \pm 0.6$, whereas the baseline method attains $51.2\% \pm 0.7$. Confidence intervals were estimated with 95% statistical significance with a *t-test*. Figure 1 shows the performance of our system for all these words. It can be seen that for paradigm classes with sizes up to 6, our method selects the correct paradigm as first option for more than 70% of the words. In addition, the percentage of cases in which the correct paradigm is between the two better scored candidates is above 90%. The only exception is the case of the paradigm class containing 32 candidate paradigms; in this case, the results are not so good due to the fact that the lexical model is harder to estimate. Results for paradigm classes of size 3 are also worse than the rest, although they are not reliable, since only two words were used to obtain the results. The information represented in the histogram shows that our method is not only useful to choose the best candidate paradigm, but also to sort the paradigm candidates in scenarios as the one depicted in the end of section 1.5.

1.8 Concluding remarks

Our previous work on enlarging monolingual dictionaries of rule-based MT systems by non-expert users has been extended with an n-gram model of lexical category and inflection information to tackle the common case of paradigm classes including more than one paradigm. Results significantly improve those of the baseline and show that the extended system can be used to successfully obtain the right paradigm for most new words; even in those cases where the inferred paradigm is wrong, our system may prove useful as it provides an ordered list of candidates which may help users validating the new entries to quickly arrive to the correct paradigm. We plan to extend our approach to other languages and explore the use of a hidden Markov model (Manning and Schütze, 1999) instead of an n-gram language model. We also plan to detect situations in which a word may be correctly added to more than one paradigm by studying the values of the perplexities of each option.

Acknowledgements

This work has been partially funded by Spanish Ministerio de Ciencia e Innovación through project TIN2009-14009-C02-01, by Generalitat Valenciana through grant ACIF/2010/174 from VALi+d programme, and by Universitat d'Alacant through project GRE11-20.

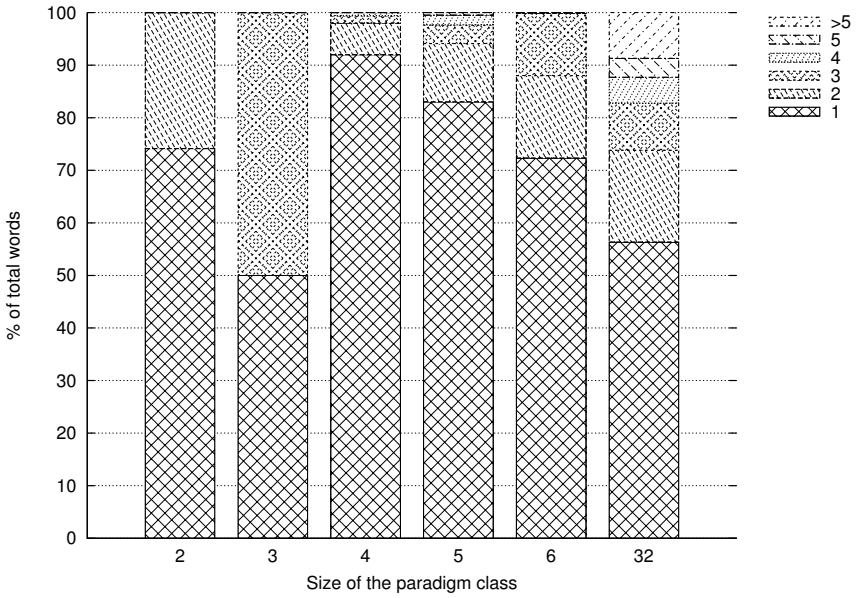

FIGURE 1 Position of the correct paradigm in the ordered list provided by our method depending on the size of the paradigm class. The number of words assignable to each class is 742, 2, 3 685, 9 289, 2 541, and 5 969, respectively. Note that only two words were found in the test set for paradigm classes with three paradigms, so results of the second bar of the histogram are not very reliable.

References

Bartusková, D. and R. Sedlácek. 2002. Tools for semi-automatic assignment of Czech nouns to declination patterns. In *Proceedings of the 5th International Conference on Text, Speech and Dialogue*, pages 159–164.

Esplà-Gomis, M., V. M. Sánchez-Cartagena, and J. A. Pérez-Ortiz. 2011. Enlarging monolingual dictionaries for machine translation with active learning and non-expert users. In *Proceedings of Recent Advances in Natural Language Processing*, pages 339–346.

Federico, M., N. Bertoldi, and M. Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Proceedings of Interspeech*.

Font-Llitjós, A. 2007. *Automatic improvement of machine translation systems*. Ph.D. thesis, Carnegie Mellon University.

Forcada, M.L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J.A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F.M. Ty-

ers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation* 25(2):127–144.

Hutchins, W. J. and H. L. Somers. 1992. *An introduction to machine translation*. Academic Press, London.

Manning, C. D. and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

McCreight, E.M. 1976. A space-economical suffix tree construction algorithm. *Journal of the Association for Computing Machinery* 23:262–272.

McShane, M., S. Nirenburg, J. Cowie, and R. Zacharski. 2002. Embedding knowledge elicitation and MT systems within a single architecture. *Machine Translation* 17:271–305.

Sánchez-Cartagena, V. M., M. Esplà-Gomis, and J. A. Pérez-Ortiz. 2012. Source-language dictionaries help non-expert users to enlarge target-language dictionaries for machine translation. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, eds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Tiedemann, J. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, vol. V, pages 237–248. John Benjamins.

Wang, A., C. D. V. Hoang, and M. Kan. 2010. Perspectives on crowdsourcing annotations for natural language processing. Tech. rep., School of Computing, National University of Singapore.