Rapid development of data for shallow transfer RBMT translation systems for highly inflective languages

Jernej Vičič

University of Primorska
Glagoljaška 8, SI-6000 Koper
jernej.vicic@upr.si.si

Abstract

The article describes a new way of constructing rule-based machine translation systems (RBMT), in particular shallow-transfer RBMT suited for related languages. The article describes methods that automate parts of the construction process. The methods were evaluated on a case study: the construction of a fully functional machine translation system of closely related language pair Slovenian - Serbian. The Slovenian language and The Serbian language belong to the group of southern Slavic languages that were spoken mostly in the former Yugoslavia. The economies of the nations where these languages are spoken are closely connected and younger generations, the post-Yugoslavia breakage generations, have difficulties in mutual communication, so there is a big interest in construction of such translation system. The system is based on Apertium (Oller and Forcada, 2006), an open-source shallow-transfer RBMT toolkit. Thorough evaluation of the translation system is presented and conclusions present the strong and the weak points of this approach and explore the grounds for further work.

1. Introduction

Slovenian language and Serbian language belong to the group of southern Slavic languages that are spoken mostly on the territory of former Yugoslavia. Slovenian language is mostly spoken in Slovenia, Serbian language is mostly spoken in Serbia. The languages share common roots and even more importantly they share common recent historical environment, these languages were spoken in the same country, even taught in schools as languages of the surroundings. Economies of both countries are closely connected. Younger generations, the post-Yugoslavia breakage generations, have difficulties in mutual communication, so there is a big interest in the construction of an automatic machine translation system for this language pair. Both languages are highly inflective and morphologically and derivationally rich languages and differ greatly from mostly used languages in electronic materials like English, Arabic, Chinese, Spanish and French. This means that most of the data and translation methods must be at least revisited or even worse rewritten. This language pair is closely related lexicographically and syntactically which simplifies most of the translation system production steps. All methods and materials discussed in this paper were tested on a fully functional machine translation system based on Apertium (Oller and Forcada, 2006; Corbi-Bellot et al., 2005), an open-source shallow-transfer RBMT toolkit. Apertium is an open-source machine translation platform, initially aimed at related-language pairs but recently expanded to deal with more divergent language pairs (such as English-Catalan). The platform provides a language-independent machine translation engine, tools to manage the linguistic data necessary to build a machine translation system for a given language pair and linguistic data for a growing number of language pairs. All these properties make Apertium a perfect choice in a cost effective machine translation system development.

The construction of a machine translation system for a new language pair falls roughly into two categories:

- A long and not particularly interesting job of manual dictionary and rule construction in case of classic Rule-Based Machine Translation (RBMT) (Hutchins, 2005) system construction approach, including all similar approaches.

- Automatic machine translation system construction in case of corpus-based machine construction systems such as Statistical Machine Translation (SMT) (Brown et al., 1993; Och and Ney, 2003) or Example-Based Machine Translation (EBMT) (Nagao, 1984) and (Hutchins, 2005). Several other examples of corpus-based machine translation systems are available.

The SMT seems like a perfect choice as some of the best performing machine translation systems are based on the SMT technologies (NIST, 2006), but it has a few drawbacks that cannot be ignored; the SMT systems, to be efficient, require huge amount of parallel text (Och, 2006) that is available only for a few of the widely used languages like English, Spanish, French, Arabic, etc. The morphologically rich and highly inflective languages like the pair presented in this paper (Slovenian and Serbian language) present an even bigger problem as shown on table 1 3. in section 3.1.. The rest of the article is organized as follows: state of the art is presented in section 2., follows a presentation of the used methods in section 3.. The evaluation methodology with results is presented in section 4., the article concludes with the discussion.

2. State of the art

According to (NIST, 2006), today's best performing automatically constructed machine translation systems , we will concentrate on SMT systems like (Google, 2008), require huge amounts of parallel data to learn from. Systems for language pairs with big parallel corpora yield good results, for some language pairs even best results overall. Such big corpora are not available for most of the languages. The smaller corpora, even linguistically annotated, are easier to be found, at least for most of the European languages, see (Dimitrova et al., 1998; Multext, 2007).

2.1. Available technologies and materials

A research of already available and accessible language processing tools and materials, mostly corpora, revealed that there is a reasonably big amount of work already done for the Slovenian language, less for the Serbian language. The tools for the Slovenian language are (reasonable or even good quality): a part of speech tagger (Erjavec, 2006; Brants, 2000), a lemmatizer (Erjavec, 2006; Erjavec, 2004), a stemmer (Popovic and Willett, 1992; Popovic and Willett, 2000). None of these tools exists for Serbian language. Both languages have solid monolingual reference corpora (going into hundreds of millions) and a small bilingual corpus (Dimitrova et al., 1998).
This research focuses mostly on the lexical level mainly for these reasons:

- The lexical level presents the starting ground for written text translation.
- The related languages, particularly the language pair we based our study upon, usually share the same sentence structure.
- Most of the translation takes place on the lexical level.
- Unlike some well-known languages, like English, southern Slavic languages express most of the meaning by inflecting words and less by word order.

Only the lexicographic modules were taken into consideration in this case study as the work on the project is still in progress. We concentrated the research on preceding modules, the lexicographic modules, as they present the basis for all translation stages. Still some basic structural transfer rules were constructed to greatly enhance translation performance at a small cost in expert hours.

3. Intention

Quite a few methods that automate some parts of the RBMT machine translation system construction have been presented and are even used as part of the construction toolkits. This article presents an attempt to automate all data creation processes of a shallow transfer machine translation system based on RBMT. The Apertium (Oller and Forcada, 2006) shallow transfer machine translation toolbox was used in our experiments although most of the methods could be applied to other systems. The data:

1. The monolingual source dictionary with morphological information for source language parsing.
2. the monolingual target dictionary with morphological information for target language generation.
3. The bilingual translation dictionary.
4. The shallow transfer rules.
5. The disambiguation data.

The monolingual dictionaries are used in shallow parsing of the source text and generation of the translation text in the target language. The bilingual dictionary is used for word-by-word translation, in our case the translation is based on lemmata. The shallow transfer rules are used to address local syntactical and morphological rules such as local word agreement and local word reordering. The morphological disambiguation of the source language morphological parsing phase was done using implicit disambiguation rules, in our case in form of the Hidden Markov Model (HMM) parameters (stochastic POS tagger), but other alternatives are possible such as methods described in (Homola and Kubon, 2008). Each item from the list was addressed by applying a known method or by introducing a new method. The methods are further presented in a separate subsection. A fully functional system was constructed using presented methods and overall performance of the whole system was evaluated.

3.1. Monolingual source and target dictionary creation

Let us look at an example from the English language; the transformation of the word walk into walked can be achieved by a morphological transformation rule (for past tense). A variation of the same rule would be used for the irregular word sleep, changing into slept. For languages that employ concatenative morphology [1] such as the majority of European languages, different forms of the same word are realized by changing the prefix and suffix of the word. Thus, slept can be derived from sleep by changing the suffix -ep to the suffix -pt. The same phenomenon, but to a much greater extent, occurs in highly inflectional languages, an example for Slovenian language is shown in table 1.

3.1.1. Paradigm creation

The words were grouped into paradigms in order to deal with multiple word-forms as Slovenian and Serbian language are both highly inflectional languages. Each paradigm is represented by:

- a typical lemma, the lemma the paradigm was constructed from
- a stem, the longest common prefix of all words in the lemma
- a set of all words split into stems and postfixes and Morpho-Syntactical Descriptors (MSDs) (Erjavec, 2004)

The annotated lexicons, lists of unique words with lemma descriptor and MSD, were extracted from corpus for both languages and paradigms were constructed using the algorithm in figure 1.

[1] words are composed of a number of morphemes concatenated together; the morphemes include the stem plus prefixes and suffixes

Table 1: All word forms for Slovenian lemma mesto (place/city)

word form	number	case
mest-o	Singular	nominative
mest-a	Singular	genitive
mest-u	Singular	dative
mest-o	Singular	accusative
mest-u	Singular	locative
mest-om	Singular	instrumental
mest-a	Plural	nominative
mest-∅	Plural	genitive
mest-om	Plural	dative
mest-a	Plural	accusative
mest-ih	Plural	locative
mest-i	Plural	instrumental
mest-i	Dual	nominative
mest-∅	Dual	genitive
mest-oma	Dual	dative
mest-i	Dual	accusative
mest-ih	Dual	locative
mest-oma	Dual	instrumental

```
//par - paradigms
for(i = 0; i < par.size; i++){
  for(j = i; j < par.size; j++){
    if(par[i].POS == par[j].POS){
      if(all entries agree){
        join(par[i], par[j])
      }
    }
  }
}
```

Figure 1: Paradigm construction algorithm

All word forms of a lemma present in the corpus are grouped into a class representing the lemma. A paradigm is constructed from each class; for each lemma. Two paradigms are joined together if lemmata of both paradigms, in the first step just two lemmata, later the number increases, have the same POS and if all entries agree: entries with same MSD have same postfix. Sets of entries of both paradigms are joined into a new set. The information about all lemmata that generated the paradigm is stored in a list enabling easy lookup. The monolingual source and target dictionaries were constructed using joined paradigms resulting in a roughly 20 times bigger lexicon than the starting.

3.2. Bilingual translation dictionary creation

The Number of word forms in a text is much bigger for highly inflective languages like the Slavic languages. Figure 4 shows the difference in number of word forms for the same corpus (Dimitrova et al., 1998) in four languages; three highly inflective Slavic languages: Slovenian, Serbian, Czech and English language as a reference.

The reduction of search space obviously increases the accuracy of the model (the word-by-word translation model). This result is not surprising, but a lot of information about the word form was lost in the process. Let us observe the phenomenon to a greater extent. The word alignment model as described in (Brown et al., 1993; Och and Ney, 2003) can be used as the basis for a new model that uses lemma+POS descriptions of the actual word forms used in the bilingual parallel corpus.

Table 2: Number of lemmata in corpus MULTEXT-EAST (Dimitrova et al., 1998)

language	number of words	lemmata
Slovenian	22134	6512
Serbian	21435	6832
Czech	23654	7263
English	11293	8182

Some simple definitions that will help the formulation of the equation 1
L - language, all words
E_L - lemmata of the language L
$E_{L(i)}$ - i^{th} lemma with all word forms

$$|L| = \sum_{i=0}^{|E_L|} E_{L(i)} \tag{1}$$

The search space is reduced from $|L|$ to $|E_L|$.
Observe the example:
Assuming that George Orwell's novel "1984", which comprises the multilingual sentence-aligned part of the (Dimitrova et al., 1998) corpus, is a good sample of a language, in our case the Slovenian language, we observe the values in figure 2 taken from table 2. The search space has been reduced from 22134 word forms to 6512 lemmata.

Original language $|L| = 22134$
Lematized language $|E_L| = 6512$

Figure 2: The reduction of the search space for the Slovenian language (small corpus MULTEXT-EAST (Dimitrova et al., 1998))

The bilingual parallel annotated corpus (Dimitrova et al., 1998) comprises original text with additional information in form of XML tags according to the TEI-P4 (Consortium, 2007) and the EAGLES (Leech and Wilson, 1996) guidelines. An example excerpt is shown on figure 3.

Each word is represented by the *lemma* (lemma of the word), *ana* (morphosyntactical description - MSD (Erjavec, 2004)) and the word form used in corpus. Only the lemma and the POS, first feature of MSD, of each word were extracted from the corpus for this task, leaving parallel sentences in lemmatised form with the POS tag. Figure 4 shows the prepared data.

An SMT word-to-word model (Brown et al., 1993; Och and Ney, 2003) was trained on the parallel, sentence aligned list extracted from the corpus, shown on figure 4. The lemmata alignment ensures much better alignment performance due to the search space reduction as described in equation 1 and in figure 2. The words from the monolingual dictionaries are aligned to the translations (bilingual lemmata pairs) through paradigms that retain the information about the included lemmata, see section 3.1.1..

```
<s id="Osl.2.3.5.11">
<w lemma="priti" ana="Vmps-dma">Prisla</w>
<w lemma="biti" ana="Vcip3d--n">sta</w>
<w lemma="do" ana="Spsg">do</w>
<w lemma="podrt" ana="Afpnsg">podrtega</w>
<w lemma="drevo" ana="Ncnsg">drevesa</w>
<c>,</c>
<w lemma="o" ana="Spsl">o</w>
<w lemma="kateri" ana="Pr-nsl----a">
katerem</w>
<w lemma="on" ana="Pp3msd--y-n">mu</w>
<w lemma="biti" ana="Vcip3s--n">je</w>
<w lemma="praviti" ana="Vmps-sfa">
pravila</w>
<c>.</c>
</s>
```

Figure 3: A sentence in the corpus

```
priti_V biti_V do_S podrt_A
drevo_N , o_S kateri_P on_P
biti_V praviti_V .
```

Figure 4: Prepared data: lemmata and POS of each word from the corpus

3.3. Transfer rules induction

This experiment focused on morphologically annotated data. The creation of shallow-transfer translation rules has been systematically avoided. A few test rules have been manually created to observe the translation quality performance boost. Shallow transfer translation rules will be automatically constructed using already available software (Sanchez-Martinez and Forcada, 2007) and automatically ordered according to (Vicic and Forcada, 2008).

3.4. Implicit disambiguation rules training

The POS tagger has been used to disambiguate source language parsing options. Two POS taggers were tested: the TnT (Brants, 2000) from TOTALE (Erjavec, 2006) toolkit and the Apertium POS tagger (Sanchez-Martinez et al., 2007). The first was already trained on the same corpus while the second was trained in an unsupervised method on an automatically harvested text from the internet. As expected, better results were achieved by (Brants, 2000) due to better training data. The experiment was not conducted thoroughly due to lack of time and due to satisfactory results achieved by the available tools.

4. Evaluation methodology and results

The evaluation of the translations was performed in four parts, each part is further described in a separate subsection in the continuation of this chapter:

1. The automatic objective evaluation using BLEU (Papineni et al., 2001) metric.
2. The automatic objective evaluation using METEOR (Banerjee and Lavie, 2005; Lavie and Agarwal, 2007) metric.
3. The non-automatic evaluation by counting the number of edits needed to produce a correct target sentence from automatically translated sentence.
4. Non-automatic subjective evaluation following (LDC, 2005) guidelines.

Subjective evaluation was performed after first poor BLEU results triggered some distrust. Many authors agree that BLEU metric systematically penalizes RBMT systems (Callison-Burch et al., 2006) and it is not suited for highly inflective languages. Authors of METEOR (Banerjee and Lavie, 2005), (Lavie and Agarwal, 2007) state that their system fixes most of the problems encountered using BLEU metric; they state that METEOR correlates highly with human judgement. Unfortunately METEOR did not support our language pair, additional software had to be written. The bilingual parallel corpus (Dimitrova et al., 1998) was used in automatic evaluation of translations. The K-fold cross-validation (Kohavi, 1995) was used as the method for estimating the generalization error as it is most suitable for small data sets. In our case five-fold cross validation was used instead of more frequently used ten-fold cross validation as construction of a fully functional system was not automated. The corpus was divided into five parts, each part consisting of roughly 1700 sentences. The evaluation consisted in selecting one part of the corpus as testing set and remaining four parts as training set. The translation system was constructed according to the methodology presented in 3. using the selected training set. The evaluated values in each fold and the average final values are presented.

4.1. Automatic objective evaluation using BLEU metric

The publicly available implementation of the BLEU metric (NIST, 2008) version v11b was used. Results are presented in table 3. These scores are relatively low, especially considering the relatedness of the language pair. Low values are partly to be attributed to high inflexibility of the language pair and partly to the fact that the BLEU metric penalizes RBMT systems (Callison-Burch et al., 2006).

Table 3: The BLEU metric scores, each fold is presented in a separate line, last two lines present average values with standard deviation

fold	BLEU value
1	0.1167
2	0.1211
3	0.1206
4	0.1198
5	0.1201
Average	0.1196
STDEV	0.0017

4.2. Automatic objective evaluation using METEOR metric

The publicly available implementation of the METEOR metric (Lavie and Agarwal, 2007) version v0.6 was used.

The METEOR uses stemming mechanism as one of the algorithms that enhance correlation between METEOR metric and human evaluation for highly inflectional languages. The stemming mechanism that is a side-product of the described translation system was used. Results are presented in table 4.

Table 4: The METEOR metric scores, each fold is presented in a separate line, last two lines present average values with standard deviation

fold	METEOR value
1	0.6344
2	0.6296
3	0.6316
4	0.6297
5	0.6352
Average	0.6321
STDEV	0.0026

4.3. Non-automatic evaluation using edit distance

The edit-distance (Levenshtein, 1965) was used to count the number of edits needed to produce a correct target sentence from automatically translated sentence. This procedure shows how much work has to be done to produce a good translation. The metric roughly reflects the complexity of post-editing task. The evaluation comprised of selecting 100 sentences from testing data, translating these sentences using the translation system and manually counting the number of words that had to be changed in order to obtain a perfect translation. By perfect translation we mean a translation that is syntactically correct and expresses the same meaning as the source sentence. 22% of all words had to be corrected in order to achieve the perfect translation. The results of this evaluation can be compared to results of the same metric used on a similar system; (Homola and Kubon, 2008). Language pair's properties and similarities of our system in comparison to (Homola and Kubon, 2008) make the comparison feasible. The (Homola and Kubon, 2008) system presents one of the best performing related language translation systems with only 3.55 % of errors and therefore presents a good reference point for our system's final goal. This evaluation was conducted as a test on a low number of test translations due to time limitations.

4.4. Non-automatic subjective evaluation following (LDC, 2005) guidelines

Subjective manual evaluation of translation quality was performed according to the annual NIST Machine Translation Evaluation Workshop by the Linguistic Data Consortium guidelines. The most widely used methodology when manually evaluating MT is to assign values from two five-point scales representing fluency and adequacy. These scales were developed for the annual NIST Machine Translation Evaluation Workshop by the Linguistic Data Consortium (LDC, 2005).
The five point scale for adequacy indicates how much of the meaning expressed in the reference translation is also expressed in a hypothesis translation:

- 5 = All
- 4 = Most
- 3 = Much
- 2 = Little
- 1 = None

The second five-point scale indicates how fluent the translation is. It expresses weather the translation is syntactically correct. When translating into Serbian the values correspond to:

- 5 = Flawless translation
- 4 = Good Serbian
- 3 = Non-native Serbian
- 2 = Disfluent Serbian
- 1 = Incomprehensible text

Separate scales for fluency and adequacy were developed under the assumption that a translation might be disfluent but contain all the information from the source. Four independent evaluators (two native speakers) evaluated sets of 100 sentences using this methodology. The results are presented in 5.

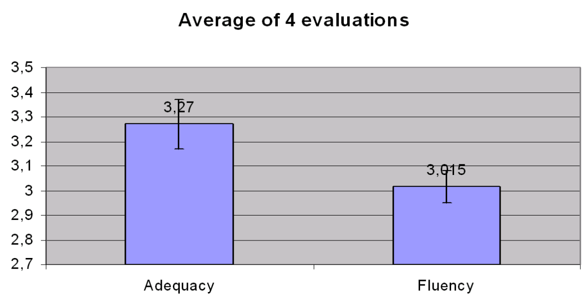

Figure 5: Evaluation results using (LDC, 2005) guidelines. Average values of four independent evaluations show high scores for adequacy and lower values for fluency.

5. discussion

The article presents an ongoing research of rapid construction of shallow-transfer machine translation systems for related languages. The evaluation shows promising results although there is still a lot of space for improvement. All described methods were tested on a fully-functional translation system, the latest version of the system is available online at the following address: http://jt.upr.si/guat/index.php.
The automatic construction of shallow-transfer translation rules has not been addressed in this research and will, in addition to automatic ordering of the rules, present the next step of the research.

6. References

S. Banerjee and A. Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*. Ann Arbor, Michigan.

Thorsten Brants. 2000. Tnt – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference*. Seattle, WA.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational linguistics*, 19(2):163–311.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *Proceedings of EACL*.

TEI Consortium. 2007. Tei p5: Guidelines for electronic text encoding and interchange. Technical report, TEI consortium.

Antonio M. Corbi-Bellot, Mikel L. Forcada, Sergio Ortiz-Rojas, Juan Antonio Prez-Ortiz, Gemma Ramirez-Snchez, Felipe Sanchez-Martnez, Inaki Alegria, Aingeru Mayor, and Kepa Sarasola. 2005. An open-source shallow-transfer machine translation engine for the romance languages of spain. In *Proceedings of the Tenth Conference of the European Association for Machine Translation*, pages 79–86, May.

Ludmila Dimitrova, Nancy Ide, Vladimir Petkevic, Tomaz Erjavec, Heiki Jaan Kaalep, and Dan Tufis. 1998. Multext-east: Parallel and comparable corpora and lexicons for six central and eastern european languages. In *COLING-ACL*, pages 315–319.

Tomaz Erjavec. 2004. Multext-east version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04*.

Tomaz Erjavec. 2006. Multilingual tokenisation, tagging, and lemmatisation with totale. In *Proceedings of the 9th INTEX/NOOJ Conference*.

Google. 2008. Google translator.

Petr Homola and Vladislav Kubon. 2008. A method of hybrid mt for related languages. In *Intelligent information systems XVI : proceedings of the International IIS '08 conference*, pages 269–278.

John Hutchins. 2005. Towards a definition of example-based machine translation. In *MT Summit X, Proceedings of Workshop on Example-Based Machine Translation*, pages 63–70.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1137–1143.

A. Lavie and A. Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of Workshop on Statistical Machine Translation at the 45th Annual Meeting of the Association of Computational Linguistics (ACL-2007)*.

LDC. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Technical report.

GN Leech and A. Wilson. 1996. Eagles recommendations for the morphosyntactic annotation of corpora. Technical report, ILC-CNR, Pisa.

V. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk*, pages 845–848.

Multext. 2007. The multext project.

Makoto Nagao. 1984. A framework of a mechanical translation between japanese and english by analogy principle. *Artificial and Human Intelligence*.

NIST. 2006. Nist 2006 machine translation evaluation official results.

NIST. 2008. Evaluation software.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29:19–51.

Franz Josef Och. 2006. Challenges in machine translation. In *Proceedings of ISCSLP*.

Carme Armentano Oller and Mikel L. Forcada. 2006. Open-source machine translation between small languages: Catalan and aranese occitan. In *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages)*, pages 51–54.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical report.

M. Popovic and P. Willett. 1992. The effectiveness of stemming for natural language access to slovene textual data. *Journal of the American Society for Information Science*, 43(5).

M. Popovic and P. Willett. 2000. Krnjenje kot osnova nekaterih nekonvencionalnih metod poizvedovanja. *Knjinica, Ljubljana*, 44.

Felipe Sanchez-Martinez and Mikel L. Forcada. 2007. Automatic induction of shallow-transfer rules for open-source machine translation. In Andy Way and Barbara Gawronska, editors, *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, volume 2007:1, pages 181–190. Skovde University Studies in Informatics, September.

Felipe Sanchez-Martinez, Carme Armentano-Oller, Juan Antonio Perez-Ortiz, and Mikel L. Forcada. 2007. Training part-of-speech taggers to build machine translation systems for less-resourced language pairs. In Vctor J. Daz Madrigal and Fernando Enrquez de Salamanca Ros, editors, *Procesamiento del Lenguaje Natural (XXIII Congreso de la Sociedad Espanola de Procesamiento del Lenguaje Natural)*, volume 39, pages 257–264, September.

Jernej Vicic and Mikel Forcada. 2008. Comparing greedy and optimal coverage strategies for shallow-transfer machine translation. In *Proceedings of Intelligent information systems*.