Pattern Recognition Letters 46 (2014) 96–103

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Statistical machine translation of subtitles for highly inflected language pair ☆

Mirjam Sepesy Maučec *, Zdravko Kačič, Darinka Verdonik

Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova 17, 2000 Maribor, Slovenia

ARTICLE INFO

Article history:
Received 12 August 2013
Available online 2 June 2014

Keywords:
Statistical machine translation
Phrase-based translation
Highly inflected languages
Bilingual dictionary
Entropy

ABSTRACT

This paper addresses the problem of statistical machine translation between highly inflected languages. Even when dealing with closely-related language pairs, statistical machine translation encounters problems if the parallel corpus is not big enough. To reduce the problem of data sparsity, we use the approach called factored translation, which has proven successful when translating between English and a morphologically rich language. We show that it is even more useful when translating between two highly inflected languages. The main contribution of the paper involves two extensions of the factored translation approach. First, we propose a new, more general asynchronous framework for training translation components, where lemmas in the lemma component and MSD tags in the MSD component are aligned independently of alignment done for surface word forms. The second contribution of the paper is a new technique for efficient use of a bilingual dictionary in the translation process. A dictionary is introduced into the lemma component to improve lexical translation. Dictionary use is based on entropy. We tested our enhanced translation approach on the Slovenian–Serbian language pair. The system was trained on a freely available OpenSubtitle corpus. The results show improvements in automatic scores (BLEU and TER). The approach could be used for other language pairs, especially if one or both are highly inflected.

1. Introduction

Language technology has brought many useful applications. Usually, most research and experiments concern widely-spoken languages [27]; however, as the META-NET White Paper Series "Languages in the European Information Society"[1] emphasise, it is crucial for the future existence of all languages that their users have the same opportunities to use modern language technology as do the users of widely-spoken languages. Even though we can find some research for less-resourced languages, these languages still have a considerable disadvantage compared to widely-spoken languages.

In this paper, we contribute to research into less-resourced languages in the domain of statistical machine translation (SMT). The usual obstacle to developing technologies for such languages is a lack of the necessary language resources. Automatic procedures can to some degree offer alternative solutions when the required tools or resources are missing. SMT has proved successful in a number of evaluations since its revival. There are many reasons for its widespread use. The most prominent one is that statistical approaches deliver very high performances. Several open-source tools are already available to train translation systems for many language pairs. When using statistical approaches, theoretically, no linguistic knowledge and very little effort are required to build a machine translation system, presuming that a parallel corpus is available. Therefore, there is a constant need for larger bilingual corpora in almost every SMT system. At present, several parallel corpora are available; however, the general issue is whether the corpus is a representative sample of the domain of interest. This paper considers the subtitle domain, using the open-source parallel corpus of subtitles, OpenSubtitles [34]. The subtitle domain was chosen because subtitling is the preferred translation method for multimedia content in Europe. Nowadays, it is also a very attractive domain for building SMT systems [23,8], because there is a clear need to optimise the productivity of current subtitle translation workflow processes.

Most research in SMT is performed on language pairs that include English. However, these experiments encounter the problem that one language of the pair is morphologically richer than the other. In our study, in contrast, we work with two highly inflectional languages: Slovenian and Serbian. Slovenian and Serbian belong to the same family of southern Slavic languages and so are closely related. Both languages feature 6 to 7 cases, multiplied

☆ This paper has been recommended for acceptance by Y. Chang.

* Corresponding author. Tel.: +386 2 220 72 25; fax: +386 2 251 11 78.
E-mail addresses: mirjam.sepesy@uni-mb.si (M.S. Maučec), zdravko.kacic@um.si (Z. Kačič), darinka.verdonik@um.si (D. Verdonik).

[1] http://www.meta-net.eu/whitepapers/overview

http://dx.doi.org/10.1016/j.patrec.2014.05.012
0167-8655/

by three genders (masculine, feminine, neuter) and singular and plural forms – and dual in the case of Slovenian. The verb forms of both languages are similarly inflectional, even though not completely parallel (e.g., Serbian has tenses such as aorist, not known in Slovenian). All this produces many different morpho-syntactic forms (in Slovenian around 20 for nouns and verbs, up to 60 and more for adjectives) of inflectional parts-of-speech, i.e. especially nouns, adjectives, verbs, pronouns and the basic numerals. For related languages, the rule-based machine translation (RBMT) approach is the most common one in use, but building a RBMT system involves considerable manual effort in order to develop the necessary resources. In our study we avoid the need for human knowledge about the languages under consideration; we rely only on available resources, although we know that these are not of desired quality.

1.1. Organization of this paper

We discuss related work in the next section. Section 3 describes the basics of phrase-based SMT. The main contribution of the paper is described in Sections 4 and 5. In Section 4 we first review the extension of phrase-based SMT into factored translation and outline the new idea of independent extraction of phrases for translation model components. We propose asynchronous training of lemma and morpho-syntactic description (MSD) components in a 2-paths-back-off system. Mathematical formulation is provided. In Section 5, more efficient introduction of a bilingual dictionary, based on entropy, is defined. Experiments are reported in Section 6. Results are evaluated by two standard automatic evaluation metrics: BLEU and TER. We also compute the ratio of exact matches and Levenshtein distance. All reported results are statistically justified. Section 7 concludes the paper.

2. Related work

Machine translation is a research field that is more than 60 years old. In the early days, many approaches were explored, ranging from simple direct translation methods to more sophisticated transfer methods. The most recent trends in machine translation are data-driven methods, especially statistical methods. The first SMT systems were based on IBM models that use words as nuclear translation units [2]. Different variations and extensions on the IBM models were defined [9,19]. Word alignments were improved by introducing maximum entropy models [12]. Translation improvements were obtained by adding linguistic information. In [29], words were lemmatized. Interpolating lemma and word alignment models further improved the results [36]. The idea of a hierarchical lexicon, where a word is represented at varying levels of inflectional specificity, was explored [24]. Data-driven morphology reduction when translating from a more-inflected language to a less-inflected language, was presented [20].

Words may not be the best atomic units for translation. Phrase-based SMT models use phrases instead of words [26,15], whereby a phrase is a contiguous multi-word sequence and has no linguistic motivation. Several methods based on phrase-based SMT have been proposed [37,13]. Linguistic knowledge can help to improve phrase-based SMT; it can be utilized during pre-processing and post-processing stages. From a search perspective, it is better to integrate these stages into one model. Factored translation models integrate linguistic annotation as factors into an extension of phrase models [16]. More complex factored models have been explored for English–Czech translation [1] and Russian–English translation [11]. The translation process can be broken down into translation and generation steps. Generating morphology when translating from a less-inflected language into a more-inflected language poses various challenges [22].

Syntax-based models were also tried, as a means of overcoming the limitations of phrase-based SMT [30]. It has been shown that, in some cases, tree-based models can outperform phrase-based models when dealing with languages with complex morphology. However, while a Slovenian parser is now available [4], the Serbian language still lacks the necessary linguistic infrastructure for syntactic parsing; therefore, this approach can not be performed on the Slovenian–Serbian language pair.

Machine translation between closely-related languages has been studied for inflectional, analytical and agglutinative languages. The Turkman–Turkish language pair was addressed in [33]. Translating a series of letters instead of words improved the results for the Catalan–Spanish language pair [35]. Shallow transfer machine translation systems have been studied for Slavic languages [10].

In this paper we examine SMT for closely-related languages in the subtitle domain. To our knowledge, no machine translation research has been performed in the subtitle domain for related languages, although subtitle translation could greatly benefit from introducing SMT. The main limitation is the lack of sufficient parallel subtitle corpora required to train the SMT models. The OpenSubtitle corpus was published recently for such research [34]. It is based on openly available subtitles with no quality checking. In this paper we will prove its usefulness for the Slovenian–Serbian language pair, though we acknowledge that it is a problematic resource for this language pair because the translations are not direct: the files for Slovenian and Serbian were translated from English. However, all small languages face this problem. The second problem we encountered was that Slovenian and Serbian are both morphologically rich languages with relatively free word order; therefore, there are many different word forms, which causes the problem of data sparsity. These complications make SMT training difficult. In this paper we propose some improvements to state-of-the-art factored phrase-based SMT by introducing a general asynchronous framework for training translation components and by using a translation dictionary.

3. Phrase-based SMT

The current best-performing SMT systems are based on phrases. The idea was originally proposed by Koehn et al. [18], and since then much research has been based on this approach. In this section we give a brief survey.

Phrase-based SMT outperforms word-based translation for many reasons: words may not be the best atomic unit for translation, because of one-to-many mappings; translating phrases instead of single words helps to resolve ambiguities. The power of phrase-based translation depends on the phrase translation table. There are many ways to acquire such a table, the most common approach being to create a word alignment between each sentence pair of the training corpus and then to extract phrase pairs that are consistent with this word alignment. The phrase-based SMT model is mathematically formulated based on the noisy-channel model:

$$\mathbf{e} = \arg\max_{\mathbf{e}} P(\mathbf{e}|\mathbf{f}) = \arg\max_{\mathbf{e}} P(\mathbf{e})P(\mathbf{f}|\mathbf{e}). \quad (1)$$

$\mathbf{f}$ is a source sentence and $\mathbf{e}$ is a target sentence. The source sentence consists of words f_j, and the target sentence of words e_i. Words f_j belong to the source vocabulary $\mathbf{F}$ and the words e_i to the target vocabulary $\mathbf{E}$. In the phrase-based model, the source sentence $\mathbf{f}$ is broken down into I phrases $\bar{f}_i$, and each source phrase $\bar{f}_i$ is translated into a target phrase $\bar{e}_i$.