Pattern Recognition 45 (2012) 3183–3192

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Explicit length modelling for statistical machine translation

Joan Albert Silvestre-Cerdà *, Jesús Andrés-Ferrer, Jorge Civera

Universitat Politècnica de València, Departament de Sistemes Informàtics i Computació, Camí de Vera s/n, 46022 València, Spain

ARTICLE INFO

Available online 21 January 2012

Keywords:
Length modelling
Log-linear models
Phrase-based models
Statistical machine translation

ABSTRACT

Explicit length modelling has been previously explored in statistical pattern recognition with successful results. In this paper, two length models along with two parameter estimation methods and two alternative parametrisations for statistical machine translation (SMT) are presented. More precisely, we incorporate explicit bilingual length modelling in a state-of-the-art log-linear SMT system as an additional feature function in order to prove the contribution of length information. Finally, a systematic evaluation on reference SMT tasks considering different language pairs proves the benefits of explicit length modelling.

1. Introduction

Explicit length modelling is a well-known problem in pattern recognition which is often disregarded. However, it has provided positive results in applications such as author recognition [25], handwritten text and speech recognition [29], and text classification [10], whenever it is taken into consideration.

Length modelling may be considered under two points of view. On the one hand, the so-called implicit modelling in which the information about the length of the sequence is indirectly captured by the model structure. This is often the case of handwritten text and speech recognition [11], language modelling [5] and machine translation [18], which often include additional states to convey length information. On the other hand, we may perform an explicit modelling by incorporating a probability distribution in the model to represent length variability in our data sample [23]. Explicit modelling can be found in language modelling [12,19], and bilingual sentence alignment and segmentation [3,9], among others.

This work focuses on explicit length modelling for statistical machine translation (SMT). The aim of SMT is to provide automatic translations between languages, based on statistical models inferred from translation examples. State-of-the-art translation systems grounded on phrase-based models implicitly model sentence length information through features, such as word and phrase penalty, that controls the number of words and phrases in the resulting translation. As discussed in more detail later, the word penalty compensates for the bias towards short sentences [4] or prevents the generation of spurious words [17], while the phrase penalty avoids the bias towards long phrases. However, in this work, we address the problem of explicit conditional length modelling at the phrase level.

State-of-the-art phrase-based systems are basically based on a large bilingual phrase dictionary, known as phrase table. Phrase tables do not model conditional phrase length correlation between corresponding phrase translations, that is, the probability of translating a source phrase made up of l words by a target phrase of m words. However, conditional phrase length models seamlessly emerge in the generative process of a bilingual phrase-based segmentation [1].

The main contribution of the current work is a systematic and extensive evaluation of explicit conditional phrase length modelling in a state-of-the-art phrase-based SMT system. To this purpose, two conditional phrase length models are proposed along with two alternative parametrisations and two different parameter estimation methods. Furthermore, strong experimental results are reported on language pairs with different degrees of relatedness.

The rest of the paper is structured as follows. The next section describes related work in SMT regarding explicit length modelling. Section 3 introduces the log-linear framework in the context of SMT and Section 4 explains the proposed conditional phrase length models. Experimental results are reported in Section 5. Finally, conclusions and future work are discussed in Section 6.

2. Related work

Explicit length modelling in SMT has received little attention since Brown's seminal paper [4] until recently. Nowadays state-of-the-art SMT systems are grounded on the paradigm of phrase-based translation [18], in which sentences are translated as segments of consecutive words. Thereby, most recent work related to explicit length modelling has been performed at the

* Corresponding author.
E-mail address: jsilvestre@dsic.upv.es (J.A. Silvestre-Cerdà).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2012.01.006

phrase level with a notable exception [27]. Explicit phrase length modelling was initially presented in [26] where the difference ratio between source and target phrase lengths is employed to phrase extraction and scoring with promising results. Zhao and Vogel [28] discussed the estimation of a phrase length model from a word fertility model [4], using this model as an additional score in their SMT system. In [8], a word-to-phrase model is proposed which includes a word-to-phrase length model. Finally, [1] describes the derivation and estimation of a phrase-to-phrase model including a model for the source and target phrase lengths.

However, none of the previous works report results on how explicit phrase length modelling contributes to the performance of a state-of-the-art phrase-based SMT system. Furthermore, phrase-length models proposed so far depend on their underlying model or phrase extraction algorithm, which differ from those employed in state-of-the-art SMT systems. The current work is inspired on the explicit phrase length model proposed in [1], but applied to a state-of-the-art phrase-based SMT system [17] and assessed on diverse language pairs in order to systematically evaluate the contribution of explicit phrase length modelling in SMT.

3. Log-linear modelling

In SMT, we formulate the problem of translating a sentence as the search of the most probable target sentence $\hat{y}$ given the source sentence x

$$\hat{y} = \arg\max_{y} \Pr(y|x). \tag{1}$$

State-of-the-art SMT systems are based on log-linear models that combine a set of feature functions to directly model this posterior probability

$$\Pr(y|x) = \frac{1}{Z(x)} \exp\left(\sum_i \lambda_i f_i(x,y)\right), \tag{2}$$

λ_i being the weight for the i-th feature function $f_i(x,y)$ and $Z(x)$ a normalisation term so that the posterior probability sums up to 1. Feature weights are usually optimised according to minimum error rate training (MERT) on a development set [20].

Conventional feature functions in phrase-based SMT systems range from those depending on word-based and phrase-based translation models [18], over that directly derived from an n-gram language model [5], to those inspired on word and phrase reordering models, and word and phrase penalties. In fact, n-gram language models and word and phrase penalties capture to some extend length information at the sentence level.

In general, n-gram language models incorporate the special end-of-sentence symbol that implicitly models sentence length information, even though it is not able to incorporate long-term constraints. This limitation produces that ill-formed sentences receive an exponentially growing probability mass depending on their length [4]. Hence, the probability of well-formed sentences exponentially decays with their length. In order to alleviate this bias towards short sentences, the word penalty feature introduces a constant bonus for each new word added to the translation. However, in phrase-based SMT systems, the word penalty avoids the generation of spurious words [17]. In any case, the word penalty feature aims at implicitly modelling sentence length information, not phrase length information, as the models proposed in this work do.

On the other hand, phrase tables suffer from a bias towards long phrases due to a similar modelling deficiency. Indeed, the phrase penalty adds a constant bonus for each additional phrase incorporated into the translation. In fact, as shown in Section 4, the phrase penalty is complementary to the proposed conditional phrase length models.

In this work, in addition to the conventional features mentioned above, additional features derived from conditional phrase length models [1] are introduced. These additional features are presented in the next section.

4. Explicit length modelling

In the phrase-based approach to SMT, the translation model considers that the source sentence x is generated by segments of consecutive words defined over the target sentence y. As in [1], in order to define these segments we introduce two hidden segmentation variables

$$p(x|y) = \sum_T \sum_{l_1^T} \sum_{m_1^T} p(x, l_1^T, m_1^T | y), \tag{3}$$

T being the number of phrases into which both sentences are to be segmented, and l_1^T and m_1^T being the source and target segmentation variables, respectively. Thus, we can factor Eq. (3) as follows:

$$p(x, l_1^T, m_1^T | y) = p(m_1^T|y)\, p(l_1^T | m_1^T, y)\, p(x | l_1^T, m_1^T, y), \tag{4}$$

where $p(m_1^T|y)$ and $p(l_1^T|m_1^T,y)$ are phrase length models, whilst $p(x|l_1^T,m_1^T,y)$ constitutes the phrase-based translation model. We can independently factorise terms in Eq. (4) from left to right,

$$p(m_1^T|y) = \prod_t p(m_t | m_1^{t-1}, y), \tag{5}$$

$$p(l_1^T|m_1^T,y) = \prod_t p(l_t | l_1^{t-1}, m_1^T, y), \tag{6}$$

$$p(x|l_1^T,m_1^T,y) = \prod_t p(x(t)|x(1),\ldots,x(t-1), l_1^T, m_1^T, y), \tag{7}$$

where t ranges over the possible segmentation positions of the target sentence, l_t and m_t are the length of the t-th source and target phrases, respectively, and $x(t)$ is the t-th source phrase.

In state-of-the-art systems, the model in Eq. (5) is approximated by the phrase penalty, which is intended to control the number of phrases involved in the construction of a translation, as previously discussed. Eq. (7) is simplified by conditioning only on the t-th target phrase to obtain the conventional phrase table, which is used as another feature,

$$p(x(t)|x(1),\ldots,x(t-1), l_1^T, m_1^T, y) \coloneqq p(x(t)|y(t)), \tag{8}$$

with parameter set, $\theta = \{p(u|v)\}$, for each source, u, and target, v, phrases. Finally, Eq. (6) is used to derive conditional phrase length models that become new feature functions of our log-linear model, and the corresponding phrase-based SMT system.

Next sections present two conditional phrase length models, namely, *standard* and *specific*, as a result of different assumptions on Eq. (6). In addition, two alternative parametrisations will be considered for each of these models, referred to as *parametric* and *non-parametric*.

4.1. Standard length models

The standard length model is derived from Eq. (6) by taking the assumption that the source length l_t only depends on the corresponding target phrase length m_t as follows:

$$p(l_t | l_1^{t-1}, m_1^T, y) \approx p(l_t | m_t). \tag{9}$$