A Principled Approach to Context-Aware Machine Translation

Rafael E. Banchs
Institute for Infocomm Research
1 Fusionopolis Way, #21-01, Singapore 138632
rembanchs@i2r.a-star.edu.sg

Abstract

This paper presents a new principled approach to context-aware machine translation. The proposed approach reformulates the posterior probability of a translation hypothesis given the source input by incorporating the source-context information as an additional conditioning variable. As a result, a new model component, which is referred to as the context-awareness model, is added into the original noisy channel framework. A specific computational implementation for the new model component is also described along with its main properties and limitations.

1 Introduction

It is well known that source-context information plays a significant role in human-based language translation (Padilla and Bajo, 1998). A similar claim can be supported for the case of Machine Translation on the grounds of the Distributional Hypothesis (Firth, 1957). According to the Distributional Hypothesis, much of the meaning of a given word is implied by its context rather than by the word itself.

In this work, we first focus our attention on the fact that the classical formulation of the statistical machine translation framework, implicitly disregards the role of source-context information within the translation generation process. Based on this, we propose a principled reformulation that allows for introducing context-awareness into the statistical machine translation framework. Then, a specific computational implementation for the newly proposed model is derived and described, along with its main properties and limitations.

The remainder of the paper is structured as follows. First, in section 2, the theoretical background and motivation for this work are presented. Then, in section 3, the proposed model derivation is described. In section 4, a specific computational implementation for the model is provided. And, finally in section 5, main conclusions and future research work are presented.

2 Theoretical Background

According to the original formulation of the translation problem within the statistical framework, the decoding process is implemented by means of a probability maximization mechanism:

$$\hat{T} = \text{argmax}_T \, p(T|S) \qquad (1)$$

which means that the most likely translation $\hat{T}$ for a source sentence S is provided by the hypothesis T that maximizes the conditional probability of T given S.

Furthermore, by considering the noisy channel approach introduced in communications theory, the formulation in (1) can be rewritten as:

$$\hat{T} = \text{argmax}_T \, p(S|T) \, p(T) \qquad (2)$$

where the likelihood $p(S|T)$ is referred to as the translation model and the prior $p(T)$ is referred to as the language model.

Notice from the resulting formulation in (2) that, as the maximization runs over the translation hypothesis space $\{T\}$, the evidence $p(S)$ is not accounted for.

This particular consequence of the mathematical representation in (2) is counterintuitive to the notion of source-context information being useful for selecting appropriate translations.

This problem becomes more relevant when the probability models in (2) are decomposed into sub-sentence level probabilities for operational purposes. Indeed, the computational implementation of (2) requires the decomposition of sentence-level probabilities $p(S|T)$ and $p(T)$ into sub-sentence level probabilities $p(s|t)$ and $p(t)$, were s and t refer to sub-sentence units, such as words or groups of words.

In the original problem formulation (Brown et al., 1993), the sentence-level translation model $p(S|T)$ in (2) is approximated by means of word-level probabilities, and the sentence-level language model $p(T)$ is approximated by means of word n-gram probabilities.

Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL 2014, pages 66–70,
Gothenburg, Sweden, April 27 2014.

Within this framework, translation probabilities at the sentence-level are estimated from word-level probabilities as follows[1]:

$$p(S|T) = \prod_k \sum_n p(s_k|t_n) \qquad (3)$$

where s_k and t_n refer to individual words occurring in S and T, respectively. The probabilities $p(s_k|t_n)$ are referred to as lexical models and they represent the probability of an individual source word s_k to be the translation of a given target word t_n. These lexical models are estimated by using word alignment probabilities.

In statistical phrase-based translation (Koehn et al., 2003), the translation model is approximated by means of phrase-level probabilities (a phrase is a bilingual pair of sub-sentence units that is consistent with the word alignments).

Within this framework, translation probabilities at the sentence-level are computed from phrase-level probabilities as follows:

$$p(S|T) = \prod_i p(s_i|t_i) \qquad (4)$$

where s_i and t_i refer to phrases (i.e. groups of words) occurring in S and T, respectively. The probabilities $p(s_i|t_i)$ are estimated by means of relative frequencies and, accordingly, they are referred to as relative frequency models.

Finally, in (Och and Ney, 2002), the maximum entropy framework was introduced into machine translation and the two-model formulation in the noisy channel approach (2) was extended to the log-linear combination of as many relevant models as can be reasonably derived from the training data. In addition, the maximum entropy framework also allows for tuning the weights in the log-linear combinations of models by means of discriminative training.

Within this framework, translation probabilities at the sentence-level are estimated from phrase-level probabilities as follows:

$$p(T|S) = \frac{1}{\zeta} \exp\{\sum_i \sum_m \lambda_m \, h_m(t_i, s_i)\} \qquad (5)$$

where $h_m(s_i, t_i)$ are referred to as feature models or functions, λ_m are the feature weights of the log-linear combination, and ζ is a normalization factor. Notice from (5) that in the maximum entropy framework the posterior probability $p(T|S)$ is modeled rather than the likelihood.

[1] For the sake of clarity additional model components such as fertility, reordering and distortion are omitted in both (3) and (4).

From (3) and (4), it is clear that source-context information is not taken into account during translation hypothesis generation. In such cases, the individual sub-sentence unit probabilities depend only on the restricted context provided by the same sub-sentence unit level as observed from the training data.

In the case of (5), on the other hand, some room is left for incorporating source-context information in the hypothesis generation process by means of context-aware feature models. This is basically done by using features that relate the occurrences of sub-sentence units with relevant source-context information of lager extension.

Several research works have already addressed the problem of incorporating source context information into the translation process within the maximum entropy framework (Carpuat and Wu, 2007; Carpuat and Wu 2008; Haque et al. 2009; España-Bonet et al. 2009; Costa-jussà and Banchs 2010; Haque et al. 2010; Banchs and Costa-jussà 2011).

In the following section, we will reformulate the translation problem, as originally described in (1), in order to provide a principled approach to context-aware machine translation for both the noisy channel and the phrase-based approaches. As seen later, this will result in the incorporation of a new model component, which can be also used as a feature function within the context of the maximum entropy framework.

3 Model Derivation

In our proposed formulation for context-aware machine translation, we assume that the most likely translation $\hat{T}$ for a source sentence S does not depends on S only, but also on the context C in which S occurs. While this information might be not too relevant when estimating probabilities at the sentence level, it certainly becomes a very useful evidence support at the sub-sentence level.

Based on this simple idea, we can reformulate the mathematical representation of the translation problem presented in (1) as follows:

$$\hat{T} = \text{argmax}_T \, p(T|S, C) \qquad (6)$$

where $p(T|S, C)$ is the conditional probability of a translation hypothesis T given the source sentence S and the context C in which S occurs. This means that the most likely translation $\hat{T}$ for a source sentence S is provided by the hypothesis T that maximizes the conditional probability of T given S and C.

For now, let us just consider the context to be any unit of source language information with larger span than the one of the units used to represent S. For instance, if S is a sentence, C can be either a paragraph or a full document; if S is a sub-sentence unit, C can be a sentence; and so on.

From the theoretical point of view, the formulation in (6) is supported by the assumptions of the Distributional Hypothesis, which states that meaning is mainly derived from context rather than from individual language units. According to this, the formulation in (6) allows for incorporating context information into the translation generation process, in a similar way humans take source-context information into account when producing a translation.

After some mathematical manipulations, the conditional probability in (6) can be rewritten as follows:

$$p(T|S,C) = \frac{p(C|S,T)\, p(S|T)\, p(T)}{p(C|S)\, p(S)} \qquad (7)$$

where $p(S|T)$ and $p(T)$ are the same translation and language model probabilities as in (2), and $p(C|S,T)$ is the conditional probability of the source-context C given the translation pair $\langle S,T\rangle$.

Notice that if the translation pair is independent of the context, i.e. $\langle S,T\rangle \perp C$, then (7) reduces to:

$$p(T|S,C) = \frac{p(S|T)\, p(T)}{p(S)} \qquad (8)$$

and the context-aware formulation in (6) reduces to the noisy channel formulation presented earlier in (2).

If we assume, on the other hand, that the translation pair is not independent of the context, the formulation in (6) can be rewritten in terms of (7) as follows:

$$\hat{T} = \text{argmax}_T\, p(C|S,T)\, p(S|T)\, p(T) \qquad (9)$$

As seen from (2) and (9), the proposed context-aware machine translation formulation is similar to the noisy channel approach formulation with the difference that a new probability model has been introduced: $p(C|S,T)$. This new model will be referred to as the context-awareness model, and it acts as a complementary model, which favors those translation hypotheses T for which the current source context C is highly probable given the translation pair $\langle S,T\rangle$.

In the same way translation probabilities $p(S|T)$ at the sentence-level can be estimated from lower-level unit probabilities, such as word or phrases, context-awareness probabilities at the sentence-level can be also estimated from lower-level unit probabilities. For instance, $p(C|S,T)$ can be approximated by means of phrase-level probabilities according to the following equation:

$$p(C|S,T) = \prod_i p(C|s_i,t_i) \qquad (10)$$

where s_i and t_i refer to phrase pairs occurring in S and T, respectively, and C is the source-context for the translation under consideration.

In the following section we develop a specific computational implementation for estimating the probabilities of the context-awareness model.

4 Model Implementation

Before developing a specific implementation for the context-awareness model in (10), we need to define what type of units s_i and t_i will be used and what kind of source-context information C will be taken into account.

Here, we will consider the phrase-based machine translation scenario, where phrase pairs $<s_i, t_i>$ are used as the building blocks of the translation generation process. Accordingly, and in order to be relevant, the span of the context information to be used must be larger than the one implicitly accounted for by the phrases.

Typically, phrases span vary from one to several words, but most of the time they remain within the sub-sentence level. Then, a context definition at the sentence-level should be appropriate for the purpose of estimating context-awareness probabilities at the phrase-level. In this way, we can consider the context evidence C to be the same sentence being translated S.

With these definitions on place, we can now propose a maximum likelihood approach for estimating context-awareness probabilities at the phrase-level. According to this, the probabilities can be computed by using relative frequencies as follows:

$$p(S|s_i,t_i) = \frac{count\,(S,s_i,t_i)}{count\,(s_i,t_i)} \qquad (11)$$

where the numerator accounts for the number of times the phrase pair $<s_i, t_i>$ has been seen along with context S in the training data, and the denominator accounts for the number of times the phrase pair $<s_i, t_i>$ has been seen along with any context in the training data.

While the computation of the denominator in (11) is trivial, i.e. it just needs to count the

number of times $<s_i, t_i>$ occurs in the parallel text, the computation of the numerator requires certain consideration.

Indeed, if we consider the context to be the source sentence being translated S, counting the number of times a phrase pair $<s_i, t_i>$ has been seen along with context S implies that S is expected to appear several times in the training data. In practice, this rarely occurs! According to this, the counts for the numerator in (11) will be zero most of the time (when the sentence being translated is not contained in the training data) or, eventually, one (when the sentence being translated is contained in the training data).

Moreover, if the sentence being translated is contained in the training data, then its translation is already known! So, why do we need to generate any translation at all?

To circumvent this apparent inconsistency of the model, and to compute proper estimates for the values of $count\ (S, s_i, t_i)$, our proposed model implementation uses fractional counts. This means that, instead of considering integer counts of exact occurrences of the context S within the training data, we will consider fractional counts to account for the occurrences of contexts that are similar to S. In order to serve this purpose, a similarity metric within the range from zero (no similarity at all) to one (maximum similarity) is required.

In this way, for each source sentence $S_{i,k}$ in the training data that is associated to the phrase pair $<s_i, t_i>$, its corresponding fractional count would be given by the similarity between $S_{i,k}$ and the input sentence being translated S.

$$fcount(S_{i,k}) = sim(S, S_{i,k}) \qquad (12)$$

According to this, the numerator in (11) can be expressed in terms of (12) as:

$$count\ (S, s_i, t_i) = \sum_k sim(S, S_{i,k}) \qquad (13)$$

and the context-awareness probability estimates can be computed as:

$$p(S|s_i, t_i) = \frac{\sum_k sim(S, S_{i,k})}{\sum_k sim(S_{i,k}, S_{i,k})} \qquad (14)$$

Notice that in (14) it is assumed that the number of times the phrase pair $<s_i, t_i>$ occurs in the parallel text, i.e. $count\ (s_i, t_i)$, is equal to the number of sentence pairs containing $<s_i, t_i>$. In other words, multiple occurrences of the same phrase pair within a bilingual sentence pair are accounted for only once.

Finally, two important differences between the context-awareness model presented here and other conventional models used in statistical machine translation must be highlighted.

First, notice that the context-awareness model is a dynamic model, in the sense that it has to be estimated at run-time. In fact, as the model probabilities depend on the input sentence to be translated, such probabilities cannot be computed beforehand as in the case of other models.

Second, different from the lexical models and relative frequencies that can be computed on both directions (source-to-target and target-to-source), a symmetric version of the context-awareness model cannot be implemented for decoding. This is basically because estimating probabilities of the form $p(T|s_i, t_i)$ requires the knowledge of the translation output T, which is not known until decoding is completed.

However, the symmetric version of the context-awareness model can be certainly used at a post-processing stage, such as in *n*-best rescoring; or, alternatively, an incremental implementation can be devised for its use during decoding.

5 Conclusions and Future Work

We have presented a new principled approach to context-aware machine translation. The proposed approach reformulates the posterior probability of a translation hypothesis given the source input by incorporating the source-context information as and additional conditioning variable. As a result, a new probability model component, the context-awareness model, has been introduced into the noisy channel approach formulation.

We also presented a specific computational implementation of the context-awareness model, in which likelihoods are estimated for the context evidence at the phrase-level based on the use of fractional counts, which can be computed by means of a similarity metric.

Future work in this area includes efficient run-time implementations and comparative evaluations of different similarity metrics to be used for computing the fractional counts. Similarly, a comparative evaluation between an incremental implementation of the symmetric version of the context-awareness model and its use in a post-processing stage should be also conducted.

Acknowledgments

The author wants to thank I^2R for its support and permission to publish this work, as well as the reviewers for their insightful comments.

References

Banchs, R.E., Costa-jussà, M. R. 2011. A Semantic Feature for Statistical Machine Translation. In Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation, ACL HLT 2011, pp. 126-134.

Brown, P., Della-Pietra, S., Della-Pietra, V., Mercer, R. 1993. The Mathematics of Statistical Machine Translation: Computational Linguistics 19(2), 263-- 311

Carpuat, M., Wu, D. 2007. How Phrase Sense Disambiguation Outperforms Word Sense Disambiguation for Statistical Machine Translation. In: 11th International Conference on Theoretical and Methodological Issues in Machine Translation. Skovde

Carpuat, M., Wu, D. 2008. Evaluation of Context-Dependent Phrasal Translation Lexicons for Statistical Machine Translation. In: 6th International Conference on Language Resources and Evaluation (LREC). Marrakech

Costa-jussà, M. R., Banchs, R.E. 2010. A Vector-Space Dynamic Feature for Phrase-Based Statistical Machine Translation. Journal of Intelligent Information Systems

España-Bonet, C., Gimenez, J., Marquez, L. 2009. Discriminative Phrase-Based Models for Arabic Machine Translation. ACM Transactions on Asian Language Information Processing Journal (Special Issue on Arabic Natural Language Processing)

Firth, J.R. 1957. A synopsis of linguistic theory 1930-1955. Studies in linguistic analysis, 51: 1-31

Haque, R., Naskar, S. K., Ma, Y., Way, A. 2009. Using Supertags as Source Language Context in SMT. In: 13th Annual Conference of the European Association for Machine Translation, pp. 234--241. Barcelona

Haque, R., Naskar, S. K., van den Bosh, A., Way, A. 2010. Supertags as Source Language Context in Hierarchical Phrase-Based SMT. In: 9th Conference of the Association for Machine Translation in the Americas (AMTA)

Koehn, P., Och, F. J., Marcu, D. 2003. Statistical Phrase-Based Translation. In: Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLTEMNLP), pp. 48--54. Edmonton

Och, F. J., Ney, H. (2002) Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In: 40th Annual Meeting of the Association for Computational Linguistics, pp. 295--302

Padilla, P., Bajo, T. (1998) Hacia un Modelo de Memoria y Atención en la Interpretación Simultánea. Quaderns: Revista de Traducció 2, 107--117