A finite-state framework for log-linear models in Machine Translation[1]

Jorge González and Francisco Casacuberta

Departamento de Sistemas Informáticos y Computación
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
{jgonzalez,fcn}@dsic.upv.es

Abstract. Log-linear models represent nowadays the state-of-the-art in statistical machine translation. There, several models are combined altogether into a whole statistical approach. Finite-state transducers constitute a special type of statistical translation model whose interest has been proved in different translation tasks. The goal of this work is to introduce a finite-state framework for a log-linear modelling approach in statistical machine translation. Results for a French-English technical translation task show the convenience of the proposed methods.

1 Introduction

Statistical machine translation is a fascinating field of *natural language processing* where models can be learned automatically from training bilingual text using efficient estimation procedures. Initial, word-based alignment models were quickly overcome by the use of the so called *phrase-based* alignment models [1]. However, the performance of machine translation systems increases significantly when additional models are used in a *log-linear modelling* approach [2]. These models are typically: phrase-based translation models, language models, sentence-length models, statistical dictionaries, etc. This approach represents the state-of-the art in statistical machine translation [3].

Given a source sentence $\mathbf{s}$, the statistical machine translation problem can be expressed by means of a log-linear approach in this manner:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \Pr(\mathbf{t}|\mathbf{s}) = \underset{\mathbf{t}}{\operatorname{argmax}} \sum_i \lambda_i f_i(\mathbf{s}, \mathbf{t})$$

Finite-state transducers constitute a special type of statistical translation model whose interest in statistical machine translation has been proved in different translation tasks [4, 5]. However, the use of additional models as in the log-linear modelling approach is not effortless in a finite-state framework.

The goal of this work is to introduce a finite-state framework for a log-linear modelling approach in statistical machine translation by means of the combination and integration of different phrase-based stochastic finite-state transducers.

[1] This work has been partially supported by the EC (FEDER) and the Spanish projects TIN2006-15694-C02-01 and the Consolider Ingenio 2010 CSD2007-00018

2 Finite state models

A stochastic finite-state transducer is a tuple $\mathcal{A} = (\Sigma, \Delta, Q, i, f, P)$, where Σ is an alphabet of input symbols, Δ is an alphabet of output symbols, Q is a finite set of states, functions $i : Q \to [0, 1]$ and $f : Q \to [0, 1]$ refer to the probability of each state to be, respectively, initial and final, and partial function $P : Q \times \{\Sigma^\star \times \Delta^\star\} \times Q \to [0, 1]$ defines a set of transitions between pairs of states in such a way that each transition is labelled with an input and an output string, and is assigned a probability. Moreover, consistency properties have to be respected for functions i, f and P in order to be able to define a distribution of probabilities on the free monoid.

2.1 Phrase-based models

The derivation of phrase-based models stems from the concept of bilingual segmentation, i.e. sequences of source words and sequences of target words. It is assumed that only segments of contiguous words are considered, the number of source segments being the number of target segments and each source segment being aligned with only one target segment and vice versa.

A monotonicity constraint would allow for the implantation of phrase-based models under a finite state framework since finite state models are restricted to be monotone. By assuming that all the possible segmentations of $\mathbf{s}$ and all the possible segmentations of $\mathbf{t}$ have not the same probability, then a language model, where the words are actually source-target phrase pairs, can be employed to estimate the probability of a monotonous bilingual segmentation.

2.2 Learning phrase-based finite-state models

A set of three different models, which are implemented by means of stochastic finite-state transducers, will be integrated into a log-linear framework. A language model for monotonous bilingual segmentations together with two translation models (a direct and an inverse model) in order to score the relations between the source and the target words in a phrase pair will be considered.

A language model of phrase-pairs. The GIATI philosophy [6, 7] has been revealed as an interesting approach to infer stochastic finite-state transducers through the modelling of languages. Rather than learning relations from a bilingual corpus, GIATI converts every sentence pair into an extended-symbol string, which is directly derived from a monotonous bilingual segmentation, in order to, straight afterwards, infer a language model from.

In fact, the ideas of GIATI, which were introduced to be applied for finite state models, are actually extensive to any other language modelling techniques. Nevertheless, in our case, the GIATI algorithm will be put into practice in a finite-state framework using finite state models such as automata and transducers. Let us describe a finite state GIATI implementation.

Given a parallel corpus consisting of a finite sample C of string pairs: first, each training pair $(\bar{x}, \bar{y}) \in \Sigma^\star \times \Delta^\star$ is transformed into a string $\bar{z} \in \Gamma^\star$ from an extended alphabet, yielding a string corpus S; then, a stochastic finite-state automaton $\mathcal{A}$ is inferred from S; finally, transition labels in $\mathcal{A}$ are turned back into pairs of strings of source/target symbols in $\Sigma^\star \times \Delta^\star$, thus converting the automaton $\mathcal{A}$ into a transducer.

Monotonous bilingual segmentations will be obtained according to some word alignments between every sentence pair in the training corpus. The alignments constrain the segmentation procedure in order to trace the smallest phrase pairs that keep the source-target word relations inside them. Every phrase pair has therefore to be considered only one symbol in a language modelling application, which will be put into practice through the estimation of n-gram backoff models together with their corresponding representation under a finite state framework.

Translation models. Phrase-based translation models have proved to provide a very natural framework for statistical machine translation. Computing the translation probability of a given phrase and hence introducing information about context, these models seem to have mostly outperformed single-word models, quickly evolving into the predominant state-of-the-art [8].

In the last years, a wide variety of techniques to generate phrase-based dictionaries have been researched and implemented. Here, however, phrase-based dictionary entries would be built from the same vocabulary set of bilingual phrase pairs as for the language model.

Phrase translation probabilities will be computed under an approximation that is based on the IBM1 model [9], using a stochastic word-based dictionary to score the translation relationships and a Poisson probability distribution to weight the phrase lengths [10]:

$$\Pr(l_s|l_t) = \frac{e^{-l_t r}(l_t r)^{l_s}}{l_s!}$$

$$\Pr(s_j^{j'}|t_i^{i'}) = \Pr(l_{s_j^{j'}}|l_{t_i^{i'}}) \prod_{j''=j}^{j'} \sum_{i''=i}^{i'} \Pr(s_{j''}|t_{i''})$$

provided that l_s and l_t are the corresponding source and target phrase lengths and $r = \bar{l}_s/\bar{l}_t$ is the ratio between their mean values. Similar equations are used for the estimation of direct translation models.

Word-based dictionaries can also be estimated by means of the GIZA++ toolkit and they are obtained at the same time that the alignments that are needed for the phrase extraction procedure are produced. This way, phrase-based translation probabilities can be pre-computed for every bilingual phrase pair and they can be represented by means of looping transitions in a finite state transducer that lacks any kind of syntax or structure. There will be then as many looping transitions in the model as the number of bilingual phrase pairs is.

3 Log-linear search

The search problem in finite state models is efficiently solved by a Viterbi-like decoding algorithm. A trellis structure is then needed to collect the best hypotheses that come from the analysis of the input throughout the three different phrase-based models.

However, since the topology of GIATI models is more complex than the one of translation models, and given that all the models share their vocabulary of bilingual phrase pairs, the trellis will be built only for the GIATI model, thus using the transition probabilities in the translation models as a way to rescore the partial hypotheses from the parsing of the input through the GIATI model.

Nevertheless, a more efficient way to handle all the models using only one data structure is to include both translation models inside the GIATI transducer.

An extended transducer can then be defined in order to take into account several functions values for the function that describes the transition set, i. e.,

$$P = Q \times \{\Sigma^{\star} \times \Delta^{\star}\} \times Q \rightarrow [0,1] \times [0,1] \times [0,1]$$

where P_i will denote the i-th function value of P. On the one hand, the original n-gram probabilities will remain in the P_0 item; on the other hand, P_1 and P_2 will represent, respectively, direct and inverse translation model probabilities, for that phrase-based transition label in the transducer.

Obviously, this model integration would cause an increment of the memory allocation since translation model probabilities would be repeated throughout the transitions of the GIATI transducer. Nevertheless, it will allow the decoding algorithm to remain nearly unalterable, thus becoming very worth the trouble.

Once a set of $\hat{\lambda}_i$ parameters have been established by means of a tuning set and a multidimensional optimization algorithm such as the *Downhill Simplex* method, an equivalent, rescored transducer with again only one function value per transition would be obtained back by computing the linear combination between the optimal weights and the corresponding set of transition probabilities. Finally, transitions with a 0-probability score can be removed from the model since they represent a transition that will never be taken, thus reducing the size of the model, then the complexity of the decoding algorithms as well.

4 Experiments

The raw multilingual text that was used for building the corpus that will be employed for the experiments was provided by the Xerox Research Center Europe (XRCE). The text contains multiple translations of several usage manuals of Xerox printers. The manuals were originally written in English and translations are available in Spanish, German and French. These texts were processed for correcting mistakes and also for obtaining different simplified versions of them. In this article, a translation application between French and English was tackled. A summary of the characteristics of this corpus is presented in Table 1.

Table 1. *Characteristics of the Xerox corpus.*

		French	English
Training	No. of sentences	52,844	
	Running words	696,124	632,737
	Vocabulary	9,898	7,787
Tuning	No. of sentences	994	
	Running words	11,779	10,832
Test	No. of sentences	984	
	Running words	11,858	11,152

4.1 Results

An optimal GIATI 4-gram transducer was built following the specifications that were introduced in section 2.2. Direct and inverse translation models were added for an integrated log-linear approach. An optimization of the λ parameters was performed using the BLEU metric as a maximization function. The optimal λ weights were 44.1% for the GIATI model, 38.4% for the direct translation model, and 17.5% for the inverse translation model. Results for the single models as well as for the best combination of λ parameters are shown in Table 2.

Table 2. *Single and combined performance of the models.*

λ weights			**metrics**		**model size**	
GIATI	direct	inverse	BLEU	WER	states	transitions
100%	0	0	33.1	56.6	786,155	1,079,618
0	100%	0	14.5	62.0	208,025	279,830
0	0	100%	14.9	78.3	208,025	279,830
44.1%	38.4%	17.5%	**34.3**	**52.3**	640,011	891,511

As expected, single translation models on their own are very poor models that need to be combined into a more general translation framework. It is remarkable, however, that a direct model does reflect a lower impact on WER than an inverse model does. Instead, a bilingual phrase-based GIATI model on its own represents an interesting model which may be considered as some sort of translation and language model at the same time. Anyway, a combination of the three models in a log-linear modelling framework clearly outperforms their individual behaviours as statistical machine translation systems.

Moreover, the resulting combined model can be smaller than the original GIATI transducer because the integrated model can take advantage of those direct or inverse translation probabilities that are 0, thus a pruning technique in order to remove all the transitions that are related to them could be applied. Not only the combined transducer presents a better translation performance but also is a 18% smaller than the original GIATI transducer, what will surely cause a significative reduction about the space and time computational requirements.

5 Conclusions

Log-linear models represent nowadays the state-of-the-art in statistical machine translation. There, several models are combined altogether into a whole statistical approach. Finite-state transducers constitute a special type of statistical translation model whose interest has been proved in different translation tasks. The goal of this work has been to introduce a phrase-based finite-state framework for a log-linear modelling approach in the field of statistical machine translation.

Several phrase-based models such as a language model of bilingual pairs, a direct translation model and an inverse translation model have been successfully integrated into a finite-state log-linear modelling approach. All the models are expressed by means of finite-state transducers and also their combination is, using a simple transducer extension. Once their optimal weights have been established, a rescored (and maybe smaller) transducer can be obtained by means of a simple linear combination. The results prove the convenience of the metodology.

References

1. Tomás, J., Casacuberta, F.: Monotone statistical translation using word groups. In: Procs. of the Machine Translation Summit VIII. (2001) 357–361
2. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29**(1) (2003) 19–51
3. Fordyce, C.S.: Overview of the IWSLT 2007 Evaluation Campaign. In: IWSLT, Trento, Italy (October 2007)
4. Casacuberta, F., Ney, H., Och, F.J., Vidal, E., Vilar, J.M., Barrachina, S., García-Varea, I., Llorens, D., Martínez, C., Molau, S.: Some approaches to statistical and finite-state speech-to-speech translation. Computer Speech & Language **18**(1) (2004) 25–47
5. González, J., Sanchis, G., Casacuberta, F.: Learning finite state transducers using bilingual phrases. In: 9th International Conference on Intelligent Text Processing and Computational Linguistics. Lecture Notes in Computer Science, Haifa, Israel (February 17 to 23 2008)
6. Casacuberta, F.: Inference of finite-state transducers by using regular grammars and morphisms. In Oliveira, A., ed.: Grammatical Inference: Algorithms and Applications. Volume 1891 of Lecture Notes in Computer Science. Springer-Verlag (2000) 1–14 5th International Colloquium Grammatical Inference -ICGI2000-. Lisboa. Portugal. Septiembre.
7. Casacuberta, F., Vidal, E., Picó, D.: Inference of finite-state transducers from regular languages. Pattern Recognition **38** (2005) 1431–1443
8. Koehn, P., Monz, C.: Manual and automatic evaluation of machine translation between european languages. In: Proceedings on the HTL-NAACL Workshop on Statistical Machine Translation, Association for Computational Linguistics (2006) 102–121
9. Brown, P.F., Pietra, S.D., Pietra, V.J.D., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics **19**(2) (1993) 263–311
10. Singh, A.K., Husain, S.: Exploring translation similarities for building a better sentence aligner. In: Proceedings of the 3rd Indian International Conference on Artificial Intelligence, Pune, India (2007)