Creating a Large-Scale Arabic to French Statistical Machine Translation System

Saša Hasan, Anas El Isbihani, Hermann Ney

Lehrstuhl für Informatik VI – Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{hasan,isbihani,ney}@cs.rwth-aachen.de

Abstract

In this work, the creation of a large-scale Arabic to French statistical machine translation system is presented. We introduce all necessary steps from corpus aquisition, preprocessing the data to training and optimizing the system and eventual evaluation. Since no corpora existed previously, we collected large amounts of data from the web. Arabic word segmentation was crucial to reduce the overall number of unknown words. We describe the phrase-based SMT system used for training and generation of the translation hypotheses. Results on the second CESTA evaluation campaign are reported. The setting was in the medical domain. The prototype reaches a favorable BLEU score of 40.8%.

1. Introduction

Statistical Machine Translation (SMT) is constantly heading towards larger translation tasks. In the last few years, the amount of available mono- and bilingual corpora has steadily increased. Today's state-of-the-art SMT systems have to cope with vast amounts of training data. Language pairs such as Chinese-English or Arabic-English are widely used in evaluations in order to establish a comparison of different systems and, thus, different approaches to statistical and data-driven machine translation. By moving from single-word translation models to phrase-based models, the computational requirements also rose immensely, which is partly compensated by utilizing more powerful machines. Clearly, the trend goes towards parallel and distributed computing.

Building a machine translation system for a new language pair from scratch is a task involving multiple steps. The advantage of data-driven systems is the possibility to build them without too much knowledge about the involved languages. It is not necessary to manually create, e.g., grammar rules that reflect diverse structural properties. Instead, data-driven systems infer all needed knowledge from large amounts of parallel texts. On the one hand, if sentence-aligned data is available, the set-up of a new language pair for a SMT system is easy and straight-forward. Word alignments, i.e. mappings from source to target words, can be trained automatically. They serve as a starting point for extraction of possible phrase pairs (cf. (Zens and Ney, 2004)). Most of the state-of-the-art SMT systems use these phrases as a starting point. On the other hand, if there are no available parallel corpora, one has to gather suitable data that can be utilized for bootstrapping parallel texts. The web is an ideal collection of a multitude of documents suitable for this task. Many web sites (e.g. in the news domain) provide articles translated in several languages.

In this paper, we will present the process of creating a statistical machine translation system for a new language pair, namely Arabic-French, which is, to our knowledge, one of the first such undertakings (cf. (Guidère, 2002)). This activity involves several steps. In Section 2, the corpus acquisition is described. Section 3 discusses the necessary preprocessing steps, i.e. tokenization, sentence alignments and Arabic word segmentation which turns out to be important to reduce the overall number of unknowns. The training of the models and hypothesis generation is presented in Section 4. The performance of the prototype, which has been evaluated in the second CESTA evaluation, is addressed in Section 5. Finally, the paper is concluded in Section 6.

2. Data acquisition

The starting point for each statistical system is the acquisition of parallel texts which are usable for training the models. The Linguistic Data Consortium (LDC) issues numerous parallel corpora aligned at sentence level for language pairs such as Arabic-English and Chinese-English (cf. projects TIDES and GALE), but a parallel bilingual corpus does not exist for Arabic-French so far.

There are several sources in the world wide web that give access to language resources both in Arabic and French, such as international organizations (e.g. Amnesty International, UN, WHO), news agencies (e.g. AFP, BBC, Reuters) and journals (e.g. Le Monde Diplomatique). Usually, the problem is to find Arabic and French articles that are translations of each other, i.e. to infer a document alignment which serves as a starting point for creating sentence-aligned bilingual corpora. In (Fry, 2005) it is reported that the use of RSS news feeds can do a favorable job for automatically gathering parallel texts for this task.

As a first prototype, we downloaded and aligned data on the document level from the archives of Amnesty International.[1] This resulted in 7.6 million running words, but with a rather limited vocabulary size of 37K. The UN documents database[2] served for gathering around 62K documents ranging from January 2001 until July 2005 with a total of 108 and 105 million running words for Arabic and French, respectively, whereas the vocabulary size increased to roughly 250K (approximately half of it being singletons). This yielded enough data to build a first prototype of a statistically-driven machine translation system for Arabic to French. Detailed corpus statistics are shown in Table 1.

[1] http://www.amnesty.org/
[2] http://documents.un.org/

		RAW DATA		PROCESSED DATA	
		ARABIC	FRENCH	ARABIC	FRENCH
TRAIN	Sentence pairs	3 499 436		4 716 101	
	Running Words	70 846 523	91 700 257	108 067 564	104 848 139
	Running Words without Punct. Marks	70 207 351	91 568 905	103 594 554	96 116 356
	Vocabulary entries	1 058 172	864 255	245 280	287 785
	Singletons	548 871	465 795	108 378	122 467
DEV	Sentence pairs	902		902	
	Running Words	19 501	22 563	30 721	25 846
	Running Words without Punct. Marks	19 430	22 302	29 355	23 383
	Vocabulary entries	6 810	4 890	3 993	3 885
	Out-Of-Vocabulary words (OOVs)	868	425	207	216
	OOVs (in vocabulary entries)	614	280	154	137
TEST	Sentences	13 134		13 134	
	Running Words	224 230	-	330 223	-
	Running Words without Punct. Marks	219 609	-	311 429	-
	Vocabulary entries	38 607	-	18 192	-
	Out-Of-Vocabulary words (OOVs)	18 365	-	8 434	-
	OOVs (in vocabulary entries)	11 394	-	4 798	-

Table 1: Corpus statistics for the UN corpus: raw data and data after preprocessing (tokenization, segmentation).

3. Corpus creation

In order to generate the parallel corpus, we used two steps to accomplish this task. First, the documents had to be aligned at sentence level. For this, a preprocessing step was carried out to tokenize the words, i.e. split punctuation marks and special characters. Secondly, an Arabic word segmentation was applied in order to reduce the overall size of the vocabulary. This step is crucial due to the rich Arabic morphology which allows for decomposing the words into several morphemes, dividing the word into a number of prefixes, the stem and eventual suffixes.

3.1. Sentence alignment

For each of the gathered documents and their corresponding translations, a sentence-aligned version had to be produced for the whole repository. We applied a method that is based on (Moore, 2002) which incorporates two steps. The first step uses a sentence-length-based model and a pruned dynamic programming search to efficiently find alignments of sentences with high probability.

The second step calculates IBM model 1 alignment probabilities and uses these word correspondences to refine the final alignment.

3.2. Arabic word segmentation

In order to reduce the number of unknowns, we tokenized the sentences and preprocessed the Arabic part by using stemming methods to split the words into prefixes, stem and suffixes, which is necessary due to the fact that some types, e.g. determiners, prepositions and pronouns, are connected to the main word. As an example, *wbAlqlm* (وبالقلم, which means "*and with the pen*") consists of a compound prefix made up of three smaller prefixes, namely *w* (و, "*and*"), *b* (ب, "*with*") and *Al* (ال, "*the*").

We use a finite-state-automaton approach that codes possible prefixes and suffixes to carry out the segmentation. With this approach, the out-of-vocabulary (OOV) rate could be reduced from 4.5% to below 1% for the development data and from 8.2% to 2.6% for the test data.

4. Training and generation

After preprocessing, the sentence-aligned data was used to train the statistical models. The word alignments are determined with the freely available GIZA++ toolkit which is widely known and utilized in the SMT community. For the translation process, we used the RWTH phrase-based system (Zens et al., 2005). Phrases are extracted from the word alignments if they are contiguous, i.e. two phrases are considered to be translations of each other if the words are aligned only within the phrase pair and not to words outside.

The basic idea of phrase-based translation is to segment a given sentence in a source language (i.e. Arabic) into phrases, then translate each of these phrases into the target language (i.e. French) and finally put them together in order to constitute the target sentence. The generation process of finding the best phrase segmentations and their corresponding translations is achieved by searching through a word graph. During search, we use a log-linear combination of several models as described in the following section.

4.1. Baseline SMT system

In statistical machine translation, we are given a source language sentence $f_1^J = f_1 \dots f_j \dots f_J$, which is to be translated into a target language sentence $e_1^I = e_1 \dots e_i \dots e_I$. Among all possible target language sentences, we will choose the sentence with the highest probability:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I | f_1^J) \right\} \quad (1)$$

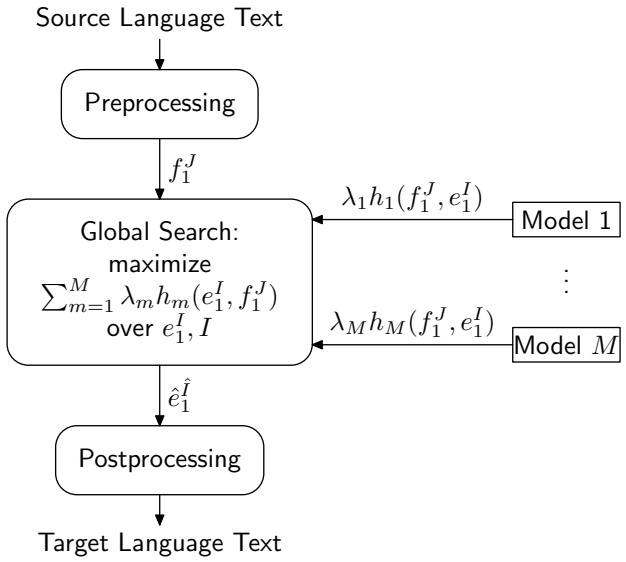

Figure 1: Overview on the direct translation model using log-linear feature combination.

The posterior probability $Pr(e_1^I|f_1^J)$ is modeled directly using a log-linear combination of several models (Och and Ney, 2002):

$$Pr(e_1^I|f_1^J) = \frac{\exp\left(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum_{I', e'^{I'}_1} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(e'^{I'}_1, f_1^J)\right)} \quad (2)$$

The denominator represents a normalization factor that depends only on the source sentence f_1^J. Therefore, we can omit it during the search process. As a decision rule, we obtain:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\} \quad (3)$$

This approach is a generalization of the source-channel approach (Brown et al., 1990). It has the advantage that additional models $h(\cdot)$ can be easily integrated into the overall system. The model scaling factors λ_1^M are trained with respect to the final translation quality measured by an error criterion (Och, 2003). An overview of the system is given in Figure 1.

We use a state-of-the-art phrase-based translation system including the following models: an n-gram language model, a phrase translation model and a word-based lexicon model. The latter two models are used for both directions: $p(f|e)$ and $p(e|f)$. Additionally, we use a word penalty and a phrase penalty. The reordering model of the baseline system is distance-based, i.e. it assigns costs based on the distance from the end position of a phrase to the start position of the next phrase.

5. Evaluation results

Our prototype has been evaluated in the second CESTA evaluation campaign for the Arabic-French track (Surcin et al., 2005). The results obtained from optimizing our system on a development set which was kindly provided by ELDA[3] look promising so far. The BLEU score of the evaluation set is 40.8% using one reference translation for automatic system evaluation. Furthermore, the evaluation data was set in the medical domain. Thus, our system was trained on extrinsic data. We expect to improve the results by incorporating more in-domain data, such as specialized medical dictionaries. From the 13 134 sentences, only a small amount was selected for final evaluation. The evaluation results of the Arabic-French prototype are given in Table 2. Some translation examples are shown in Table 3. For detailed results, see (Choukri et al., 2006).

BLEU[%]	NIST	WER[%]	PER[%]
40.84	8.94	54.8	42.5

Table 2: Case-sensitive evaluation results in the medical domain for the Arabic-French prototype based on UN data. Note that these results are based on a reduced test set from the one denoted in Table 1.

5.1. Error analysis

A brief error analysis of the translation output shows typical problems when dealing with Arabic translation:

- The form of the verb and its context determine a missing pronoun implicitly in Arabic (cf. pro-drop languages). A common case of this error is a translation such as "*Peut leur expliquer comment les types ...*" with a missing subject. A correct translation would be "*On peut leur expliquer ...*".
- The absence of copulative verbs, e.g. *être*, is frequently observed in Arabic. Thus, translations with missing copulas such as "*pratiquement à éradiquer la polio*" are common. A correct translation would be "*la polio est en voie d'éradication*".
- Due to the rich Arabic morphology, the translated verb forms are often incorrect.
- VSO word order (Verb-Subject-Object) is common in Arabic. Thus, some translations have a wrong word order, e.g. "*ont été environ 76 000 personnes*" instead of "*près de 76 000 personnes sont devenues*".

6. Conclusion

In this paper, we presented one of the first statistically-driven machine translation systems for Arabic to French. It will be applied in an open domain task with large vocabulary. We described the necessary steps to create such a system: corpus acquisition, preprocessing (such as Arabic word segmentation), training the models and finally generating translations. The prototype has been evaluated in the second CESTA campaign on data from the medical domain and the results look promising so far. Further steps are planned for the near future, such as gathering even more data from additional sources (such as WHO or news agencies) and applying more complex translation models, e.g. by using phrase reordering techniques.

[3] `http://www.elda.org/`

Arabic source sentence	*Generated French translation*
الأطفال أقل من ٦ أشهر يجب أن يكونوا تحت إشراف الطبيب عند الإصابة بالحمى.	Les enfants de moins de 6 mois doivent être sous la supervision du médecin en cas de fièvre.
فإن طرق التشخيص و العلاج تختلف من شخص لآخر حسب ظروف كل فرد.	Les méthodes de diagnostic et de traitement varie d'une personne à l'autre selon les circonstances.
هذه هي بعض الأسباب التي جعلت اليونيسف تُركِّز الاهتمام على تعليم الفتيات. أما السبب الآخر فهو أن تسريع سير العمل في تعليم الفتيات يمكن أن يشجِّع ويجمِّع المكاسب في مجالات الأولويات الأخرى لدى اليونيسف بما في ذلك محاربة فيروس نقص المناعة البشرية المكتسب وحماية الأطفال من الإساءة والاستغلال، والتشجيع على التحصين ضد الأمراض، وضمان حق الطفل في البقاء على قيد الحياة والنمو بقوة.	Telles sont les raisons pour lesquelles l'UNICEF met l'accent sur l'éducation des filles. Par ailleurs peuvent accélérer les progrès dans l'éducation des filles peut favoriser les progrès dans les domaines des autres priorités de l'UNICEF, y compris la lutte contre le VIH/SIDA et la protection des enfants contre les sévices, l'exploitation et les encourager à la vaccination contre les maladies, et garantir le droit de l'enfant à la survie, la croissance.
وبعض الأنواع الأخرى من الكتاراكت في المرضى الأصغر سنا المصابين بمرض السكر تتكون بسرعة خلال شهور وقد تسبب تدهور القدرة على الإبصار .	Et d'autres types d'UNKOWN_الكتاراكت dans des jeunes personnes souffrant du diabète est rapidement pendant les mois a causé une détérioration de la capacité de vision.

Table 3: Translation examples for the Arabic-French translation system.

7. Acknowledgment

This work has been partly supported by the R&D project TRAMES managed by Bertin Technologies as prime contractor and operated by the french DGA (Délégation Générale pour l'Armement).

8. References

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, June.

Khalid Choukri, Marianne Dabbadie, Olivier Hamon, Antony Hartley, Widad Mustafa El Hadi, Andrei Popescu-Belis, Martin Rajman, Sylvain Surcin, and Ismaïl Timimi. 2006. CESTA machine translation evaluation campaign: towards a reliable, reusable protocol. In *Proc. of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, May.

John Fry. 2005. Assembling a parallel corpus from RSS news feeds. In *Proc. of the Workshop on Example-Based Machine Translation, MT Summit X*, Phuket, Thailand, September.

Mathieu Guidère. 2002. Toward corpus-based machine translation for standard Arabic. *Translation Journal*, 6(1), January.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Machine Translation: From Research to Real Users, Proc. of the Fifth Conference of the Association for Machine Translation in the Americas*, pages 135–244, Tiburon, CA. Springer-Verlag, Heidelberg, Germany.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, July.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

Sylvain Surcin, Olivier Hamon, Antony Hartley, Martin Rajman, Andrei Popescu-Belis, Widad Mustafa El Hadi, Ismaïl Timimi, Marianne Dabbadie, and Khalid Choukri. 2005. Evaluation of machine translation with predictive metrics beyond BLEU/NIST: CESTA evaluation campaign #1. In *Proc. of the MT Summit X*, Phuket, Thailand, September.

Richard Zens and Hermann Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proc. of the Human Language Technology Conf. (HLT-NAACL)*, pages 257–264, Boston, MA, May.

Richard Zens, Oliver Bender, Saša Hasan, Shahram Khadivi, Evgeny Matusov, Jia Xu, Yuqi Zhang, and Hermann Ney. 2005. The RWTH phrase-based statistical machine translation system. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 155–162, Pittsburgh, PA, October.