A multi-genre SMT system for Arabic to French

Saša Hasan and Hermann Ney

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6 – Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{hasan,ney}@cs.rwth-aachen.de

Abstract

This work presents improvements of a large-scale Arabic to French statistical machine translation system over a period of three years. The development includes better preprocessing, more training data, additional genre-specific tuning for different domains, namely newswire text and broadcast news transcripts, and improved domain-dependent language models. Starting with an early prototype in 2005 that participated in the second CESTA evaluation, the system was further upgraded to achieve favorable BLEU scores of 44.8% for the text and 41.1% for the audio setting. These results are compared to a system based on the freely available Moses toolkit. We show significant gains both in terms of translation quality (up to +1.2% BLEU absolute) and translation speed (up to 16 times faster) for comparable configuration settings.

1. Introduction

This paper presents recent improvements of a large-scale statistical machine translation (SMT) system for the Arabic-French language pair (Hasan et al., 2006) as being researched in the TRAMES[1] project. This system is one of the first for Arabic to French that is statistically trained on a large amount of bilingual data which was gathered for this purpose, whereas early systems for this language pair, e.g. (Mankai and Mili, 1995) or (Alsharaf et al., 2004) (although for the reverse direction French-Arabic) were rule-based and lack detailed evaluation of translation quality.

Among the most important upgrades to the 2005 prototype, we report on using a larger training corpus incorporating additional data, applying improved preprocessing and general fine tuning of the system for several settings, i.e. text translation (e.g. news articles) and audio transcripts (e.g. broadcast news (BN)). The system is tuned for online translation capabilities resulting in speeds of up to 250 words per second. Memory efficiency is obtained by using a binary format of phrase tables with load-on-demand capabilities. Compared to the first prototype published at LREC'06 which serves as a baseline, the updated system achieves significant improvements of around +4% BLEU for the text setting and a favorable +20% BLEU for the audio setting.

Recent trends in SMT move away from single-purpose systems and focus on multi-genre systems instead, capable of dealing with more than one type of input text, cf. e.g. latest GALE or NIST evaluations where the tracks are split into newswire, web texts, broadcast news and broadcast conversation genres. In addition to translating written text, in particular news as well as web texts such as newsgroup articles or weblog entries, much research has been devoted to systems providing speech translation capabilities, i.e. translating audio transcripts such as broadcast news and conversations from TV or radio networks. In an ideal setting, these hybrid systems detect the type of input provided to them and actuate the corresponding settings that maximize performance for each genre. In the following, we present a system that is able to deal with several input types which are manually defined by the user so far, although we see no restriction of applying text classification methods to determine the genre on-the-fly.

In Section 2, we present the system and give details on the decoder used for generating translation hypotheses. Some room is devoted to the process of creating a bilingual training corpus, since no data was initially available for the Arabic-French language pair. Much of the overall progress, especially for translating broadcast news transcripts (i.e. audio setting), was due to genre-specific tuning of the system which is described in Section 3. The experimental setting and a discussion of the results is given in Section 4. In order to show that the obtained results are state-of-the-art, we compare the system to a widely used open source toolkit for SMT, Moses. Concluding remarks can be found in Section 5.

2. System core

The predominant approach for machine translation nowadays is data-driven and phrase-based. The building blocks are source-target phrase pairs extracted automatically from a large amount of bilingual training data, see e.g. (Zens et al., 2002; Koehn et al., 2003). Most approaches are based on a log-linear framework that combines several models in search and tries to find the best translation hypotheses by tuning the system parameters on held-out development data in order to maximize translation accuracy with respect to a set of reference translations (Och, 2003).

The system core used in this work is a phrase-based statistical machine translation system as presented in (Bender et al., 2007). The incorporated models used during decoding are a phrase-based and word-based lexicon model in both translation directions, a language model, phrase count features, length-based models for word and phrase penalty, as well as a reordering (or distortion) model based on the jump width. Since the system is designed for online translation speeds, the second translation pass using rescoring of n-best lists is omitted. The model scaling factors are tuned discriminatively in order to maximize BLEU scores on separate development sets for the different genres. We achieved significant improvements for the audio setting by

[1] Traduction Automatique par Méthodes Statistiques

	2005 system		2007 system	
	Arabic	French	Arabic	French
Doc. pairs	62K		74K	
Sent. pairs	4.7M		6.6M	
Run. words	108.1M	104.8M	151.3M	180.2M
Vocabulary	245K	288K	427K	301K

Table 1: Comparison of corpus sizes of the 2005 prototype and the final upgrade in 2007.

	Text setting		Audio setting	
	Arabic	French	Arabic	French
Doc. pairs	30		7	
Sentences	824	3 296 (4x)	466	1 864 (4x)
Run. words	22 045	102 087	16 847	91 557
Vocabulary	4 441	6 335	5 952	6 943
OOV rate	0.40%	-	1.1%	-

Table 2: Test data for the text (CESTA run2 evaluation data) and audio (Arabic broadcast news) setting.

retuning the system's parameters on a genre-specific BN development set. The pruning parameters of the decoder affect its beam size and can be used to control the speed and quality of the system output. Low beam sizes (i.e. high amount of hypotheses pruning) lessens the translation quality to some extent but obtains high throughput in terms of the number of words that are translated per second.

2.1. Training data

The first system prototype mainly incorporated data gathered from the UN documents database for the period from 2001 to 2005, totaling in 62K documents. In the latest system update, we added recently published documents up to April 2007, the archives of Amnesty International and articles from Le Monde Diplomatique.

For the audio track, broadcast news transcripts of Arabic TV and radio stations (e.g. Orient, Qatar, BBC, Alarabiya, Aljazeera, Alalam) were produced for the TRAMES project by Vecsys[2]. In total, approximately 250 audio documents consisting of 90 hours radio and TV broadcasts were transcribed resulting in 21K sentences with 585K running words of domain-specific material for the audio domain. This corpus significantly improved performance on the audio setting (cf. Section 4.).

Detailed statistics of the training corpora can be seen in Table 1. The size of the source language vocabulary increased significantly which is due to a modified preprocessing scheme which does not overly segment the Arabic words. All data was merged and used for joint phrase tables and language models. The various settings can be controlled using different configurations for the scaling factors of the log-linear models which will be explained in more detail in Section 3.

2.2. Improved preprocessing

The initial preprocessing of the baseline system used a frequency-based approach that determined segmentation points of Arabic words given a set of prefixes and suffixes and their corresponding frequencies. The segmentation for the updated system was further refined by (El Isbihani et al., 2006) and incorporates a finite state automaton-based approach by splitting compound words depending on the context already split so far. It could be shown that this approach outperforms the simple frequency-based approach which tends to segment words too excessively and results in a small vocabulary size but less translation accuracy. The finite state-based method reverses this effect. Due to the different approaches, the vocabulary sizes and number of running words vary significantly, as can be seen in Table 1.

[2] http://www.vecsys.fr/

3. Genre-specific tuning

The engine core is a phrase-based translation system using a log-linear interpolation of several models (such as phrase translation model, word-based lexicon model and n-gram language model) that determine the quality of the translation hypotheses during generation. A beam search is applied to find the best translation candidates. Additionally, the system is able to produce n-best candidates extracted from a word graph which can be further processed and reranked. In the primary mode of operation, i.e. producing translations preferably in real-time, this step is omitted and single-best translations are used. The parameters are tuned on held-out data (development set) using Maximum BLEU training by the Downhill Simplex algorithm.

The 2005 system participated in the second CESTA evaluation held in October 2005 as only participant for the Arabic-French track and scored a favorable BLEU score of 40.8% (case-sensitive). The evaluation data originated in the medical domain, i.e. news articles from the web site of the World Health Organization (WHO). The effect of adaptation to this domain is reported in (Hamon et al., 2007).

As a second stage of operability, a mode for translating audio transcrips was added to the system. For this purpose, an additional corpus has been produced for the TRAMES project (cf. Section 2.1.) and was added to the phrase table and language model training procedures. The next section shows large improvements based on this step for the audio setting. An important aspect was genre-specific tuning of the system's parameters, cf. also (Bender et al., 2007). The adjusted model scaling factors help to adapt the system from the text to the audio domain and achieve a significant improvement in BLEU score.

The parameters of the system can be set on-the-fly which enables the translation server to be started once and operate according to the user's preferences who manually defines what kind of input data is used. In future extensions, one could apply text classification methods that do this step automatically and let the system adapt to the determined domain without the user's input.

4. Experimental results

This section presents translation results for the two test sets used in the TRAMES project. Table 2 summarizes the two test data conditions. For the text setting, the official CESTA run2 evaluation data is used, whereas for the audio setting, a special subset of the BN transcriptions is chosen. Each

	1st sys 2005	2nd sys 2006	+BN-LM	3rd sys 2007
CESTA run2	40.8	42.9	43.8	**44.8**
Arabic BN				
text setting	20.9	29.7	-	34.4
audio setting	-	34.4	37.6	**41.1**

Table 3: Results of the system updates on various test settings (CESTA run2 text, BN audio transcripts, 4 reference translations each, case-sensitive BLEU, scores in percent).

evaluation corpus contains 4 reference translations on the target language side.

The performance of the system over time is shown in Table 3. The "starting" point is the 2005 prototype system trained on documents of the UN as reported in (Hasan et al., 2006). It participated in the second CESTA evaluation campaign and achieved a BLEU score of 40.8% (case-sensitive evaluation). The second system update incorporated the improved preprocessing, manual BN transcripts, additional data like Amnesty International and Le Monde Diplomatique and was tuned separately for text and audio condition. As can be seen, the genre-specific tuning results in significant performance gain on the audio setting, i.e. 34.4% instead of 29.7%, whereas the additional BN transcripts boost overall system performance from 20.9% to 29.7%, which is a 42% relative improvement. A specially tuned language model on 700M running words of additional French data (newspapers and newswire, additional audio transcripts, web data) provided by LIMSI increased performance for another 3.2% absolute.

Finally, downloading and incorporating a large number of additional documents of the UN database and retuning system parameters incorporating the additional BN-LM resulted in the third system upgrade which increased overall system performance for both settings, ending in 44.8% BLEU for text and 41.1% for audio, respectively. There is a difference of 6.7% BLEU between the text and audio setting on the Arabic BN test set, roughly half of it being due to a high BLEU brevity penalty of 0.91 (BLEU precision is 38.0%), which shows the importance of genre-specific tuning of the system's parameters.

The configurations were chosen in terms of best tradeoff between quality and speed. For the results reported, we use a 5-gram language model on the text data and an interpolated one including the 4-gram broadcast news LM on the audio data. This setting achieves translation speeds between 40 words/sec (audio) and 100 words/sec (text). When using a 4-gram LM and a slightly smaller beam size, the quality drops down to 43.4% and 40.0% BLEU for text and audio, respectively, but boosts translation speed up to 250 words per second (cf. next section).

4.1. Comparison to Moses

We compare the final system to Moses[3] (Koehn et al., 2007), an open-source translation toolkit. The Moses system uses its own implementation of phrase extraction, phrase scoring (although similar to the approach used in the TRAMES system, i.e. based on relative frequencies) and minimum error rate training. The results are shown in Table 4 for both CESTA run2 and Arabic BN test data.

[3]available at `http://www.statmt.org/moses`

	BLEU [%]	TER [%]	Translation speed [words/sec]
CESTA run2			
Moses	42.2	52.25	14.2
TRAMES	43.4	51.30	222.0
Arabic BN			
Moses	39.5	53.37	18.6
TRAMES	40.0	52.93	249.3

Table 4: Translation results and speeds for Moses and the TRAMES system. Phrase extraction is carried out on the same word alignments, same 4-gram language model is used, comparable beam sizes.

The systems use the same word alignments as starting points for phrase extraction and scoring and incorporate identical 4-gram LMs for text and audio. The beam size is adjusted to reflect the same amount of pruning. As can be seen, the performance of the TRAMES system is slightly better for both text (+1.2%) and audio (+0.5%) which might partly be due to the additional phrase count features which are missing in the Moses decoder. What stands out more are the translation speeds. The TRAMES system is around 16 times faster than Moses in the text domain, resulting in 222 words/sec, whereas audio is translated 13 times as fast as in Moses, resulting in roughly 250 words/sec. All experiments were carried out on a 2.2 GHz AMD Opteron and used less than 3.5 GB of memory.

4.2. Translation examples

Translation examples are shown in Table 5 for the text and Table 6 for the audio setting. Increasing translation quality can be noted when comparing the three main system outputs (in 2005, 2006 and 2007) to the reference translation. As can be seen, the non-adapted first system has problems with the audio domain: the OOV rate here is 7.8%, resulting in many unknowns (which are marked with label "UNK_" in the text). The BN data upgrade in 2006 fixed this problem.

5. Conclusion

In this paper, we presented an up-to-date statistically-driven machine translation system for Arabic to French that is capable of producing high-quality genre-specific translations for text and audio domains in real-time. The progress over the years was achieved using more and particularly genre-specific training data, tuning the system extensively for the various settings, incorporating better Arabic preprocessing and optimizing the decoder for maximum throughput.

The overall improvement over a three-year period was from 40.8% to 44.8% BLEU for text input and from 20.9% to 41.1% for audio transcripts. The remarkable gain for audio is mostly due to additional domain-specific training data, i.e. Arabic broadcast news transcripts, that were added to the system during the second upgrade and retuning the parameters for this setting. An additional in-domain language model (BN-LM) and more data gathered in 2007 advance

Arabic source	ويتم التركيز على الوقاية من انتقال هذا المرض من الأم إلى الطفل واتخاذ نهج للنهوض بالوعي العام بين الشباب.
French sys1 2005	et met l'accent sur la prévention de cette maladie de la mère à l'enfant et une démarche pour la promotion de la sensibilisation du public chez les jeunes.
French sys2 2006	L'accent est mis sur la prévention de la transmission de la mère à l'enfant et une approche pour la promotion de la sensibilisation du public chez les jeunes.
French sys3 2007	L'accent est mis sur la prévention de la transmission de la maladie de la mère à l'enfant et une approche pour promouvoir une prise de conscience parmi les jeunes.
French reference translation	L'accent est mis sur la prévention de la transmission de cette maladie de la mère à l'enfant et l'adoption de la démarche de la généralisation de la prise de conscience parmi les jeunes.

Table 5: Translation example showing overall progress on CESTA run2 test data (text setting) after different system upgrades over a three-year period.

Arabic source	رياض محمد رصد ردود الشارع الإيراني حيال محاكمة صدام ووافانا بالتقرير التالي.
French sys1 2005	Riyad Mohammed suivi réponses la rue des UNK_إيراني pour juger Saddam et UNK_وافانا du rapport UNK_التالي.
French sys2 2006	Riad Mohamad de suivre les mesures prises par la rue iranienne par juger Saddam et nous a fait parvenir le rapport suivant.
French sys3 2007	Riad Mohamad suivi de la réponse de la rue iranienne envers le procès de Saddam et nous a fait parvenir le rapport suivant.
French reference translation	Riad Mohamed a scruté les réactions dans la rue iranienne au sujet du procès de Saddam et nous a préparé le reportage suivant.

Table 6: Translation example showing overall progress on Arabic broadcast news test data (audio setting) after different system upgrades over a three-year period.

overall system performance to a state-of-the-art translation engine. Furthermore, the system was designed for memory efficiency such that it can be used on standard laptops. This was made possible by using a binary phrase-table representation with load-on-demand capabilities from disk. Overall memory consumption for high-quality output is below 3 GBs and translation speeds of up to 250 words per second are achieved with only a small loss in translation quality.

Acknowledgements

The authors would like to thank Gilles Adda from LIMSI-CNRS for the broadcast news language model.

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).

6. References

H. Alsharaf, S. Cardey, and P. Greenfield. 2004. French to Arabic machine translation: the specificity of language couples. In *Proc. of the 9th Annual Workshop of the European Association for Machine Translation (EAMT)*, Malta, April.

O. Bender, E. Matusov, S. Hahn, S. Hasan, S. Khadivi, and H. Ney. 2007. The RWTH Arabic-to-English spoken language translation system. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pages 396–401, Kyoto, Japan, December.

A. El Isbihani, S. Khadivi, O. Bender, and H. Ney. 2006. Morpho-syntactic Arabic preprocessing for Arabic to English statistical machine translation. In *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL), Workshop on Statistical Machine Translation*, pages 15–22, New York City, June.

O. Hamon, A. Hartley, A. Popescu-Belis, and K. Choukri. 2007. Assessing human and automated quality judgments in the French MT evaluation campaign CESTA. In *Proc. of the Machine Translation Summit XI*, Copenhagen, Denmark, September.

S. Hasan, A. El Isbihani, and H. Ney. 2006. Creating a large-scale Arabic to French statistical machine translation system. In *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*, pages 855–858, Genoa, Italy, May.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. of the Human Language Technology Conf. (HLT-NAACL)*, pages 127–133, Edmonton, Canada, May/June.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic, June.

C. Mankai and A. Mili. 1995. Machine translation from Arabic to English and French. *Information Sciences*, 3(2):91–109, March.

F. J. 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.

R. Zens, F. J. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In M. Jarke, J. Koehler, and G. Lakemeyer, editors, *25th German Conf. on Artificial Intelligence (KI2002)*, volume 2479 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 18–32, Aachen, Germany, September. Springer Verlag.