Benchmarking SMT Performance for Farsi Using the TEP++ Corpus

Peyman Passban, Andy Way, Qun Liu
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland

{ppassban,away,qliu}@computing.dcu.ie

Abstract

Statistical machine translation (SMT) suffers from various problems which are exacerbated where training data is in short supply. In this paper we address the data sparsity problem in the Farsi (Persian) language and introduce a new parallel corpus, TEP++. Compared to previous results the new dataset is more efficient for Farsi SMT engines and yields better output. In our experiments using TEP++ as bilingual training data and BLEU as a metric, we achieved improvements of +11.17 (60%) and +7.76 (63.92%) in the Farsi–English and English–Farsi directions, respectively. Furthermore we describe an engine (SF2FF) to translate between formal and informal Farsi which in terms of syntax and terminology can be seen as different languages. The SF2FF engine also works as an intelligent normalizer for Farsi texts. To demonstrate its use, SF2FF was used to clean the IWSLT–2013 dataset to produce normalized data, which gave improvements in translation quality over FBK's Farsi engine when used as training data.

1 Introduction

In SMT (Koehn et al., 2003), where the bilingual knowledge comes from parallel corpora, having large datasets is crucial. This issue is compounded when working with low-resource languages, such as Farsi. The poor performance of existing systems for the Farsi–English pair confirms the necessity of developing a large and representative dataset. Clearly all the existing problems do not originate solely from the data, but not having a reliable training set prevents us from investigating Farsi SMT to the best extent possible.

Generating datasets is a time-consuming and expensive process, especially for SMT, in which massive amount of aligned bilingual sentences are required. Accordingly instead of starting from scratch we enriched and refined the existing corpus TEP (Pilevar et al., 2011).[1] Despite having a larger alternative (the Mizan[2] corpus), TEP was selected as the basis of our work that we clarify further in Section 3 and 4.1. TEP is a collection of film subtitles in spoken/informal Farsi (SF) that have distinct structures from formal/journalistic Farsi (FF). Accordingly, training an MT engine using this type of data might provide unsatisfactory results when working with FF which is the dominant language of Farsi texts. For this reason TEP was firstly refined both manually and automatically, which Section 3 explains in detail. TEP++ is the refined version of TEP that is much closer to FF and considerably cleaner. Using both TEP and TEP++ we trained several engines for bidirectional translation of the Farsi–English pair, as well as an engine to translate between FF and SF (SF2FF). The next sections explain the challenges of dealing with SF and describe the data preparation process in detail. The structure of paper is as follows. Section 2 discusses background of MT, addressing existing systems (§2.1) and available corpora (§2.2). Section 3 explains TEP++ and our development process. Experimental results are reported in Section 4 in-

[1] TEP: Tehran English-Persian parallel corpus
http://opus.lingfil.uu.se/TEP.php
[2] http://www.dadegan.ir/catalog/mizan

cluding a comparison of the various MT systems and a study of the impact of SF2FF in Farsi SMT. Finally the last section concludes the paper along with some avenues for future works.

2 Background

Building an SMT engine for Farsi is difficult due to its rich morphology and inconsistent orthography (Rasooli et al., 2013). Not only these challenges but also the complex syntax and several exceptional rules in the grammar make the process considerably complex. The lack of data is another obstacle in this field. Nevertheless there have been some previous attempts at Farsi SMT. In this section we briefly review previous works encompassing systems in the first section, as well as available resources in the second section.

2.1 Farsi MT Systems

There are a limited number of SMT systems for Farsi. Some instances translate in one direction and some others are working bidirectionally. The Pars translator[3] is a commercial rule-based engine for English–Farsi translation. It contains 1.5 million words in its database and includes specific dictionaries for 33 different fields of science. Another English–Farsi MT system was developed by the Iran Supreme Council of Information.[4] Postchi[5] is a bidirectional system listed among the EuroMatrix[6] systems for the Farsi language. These systems are not terribly robust or precise examples of Farsi SMT and are usually the by-products of research or commercial projects. The only system that has officially been reported for the purpose of Farsi SMT is FBK's system (Bertoldi et al., 2013). It was tested on a publicly available dataset and from this viewpoint is the most important system for our purposes.[7]

2.2 Parallel Corpora for Farsi SMT

The first attempts at generating Farsi–English parallel corpora are documented in the Shiraz project (Zajac et al., 2000). The authors constructed a corpus of 3000 parallel sentences, which were translated manually from monolingual online Farsi documents at New Mexico State University. More recently Qasemizadeh et al. (2007) participated in the Farsi part of MULTEXT-EAST[8] project (Erjavec, 2010) and developed about 6000 sentences. There is also a corpus available in ELRA[9] consisting of about 3,500,000 English and Farsi words aligned at sentence level (about 100,000 sentences). This is a mixed domain dataset including a variety of text types such as art, law, culture, literature, poetry, proverbs, religion etc. PEN (Parallel English–Persian News corpus) is another small corpus (Farajian, 2011) generated semi-automatically. It includes almost 30,000 sentences. Farajian developed a method to find similar sentence pairs and for quality assurance used Google Translate.[10] All these corpora are relatively small-scale datasets. However, there are two other large-scale collections, namely Mizan and TEP, that are more interesting for our purposes. Mizan is a bilingual Farsi–English corpus of more than one million aligned sentences, which was developed by the Dadegan research group.[11] Sentences are gathered from classical literature with an average length of 15 words each. Despite comprising a large amount of sentences, the results obtained from using Mizan as a training set are less satisfactory. We will discuss the structure of Mizan and analyse some translation errors that ensue in the next section. The final corpus that is the basis of our work is TEP (Pilevar et al., 2011), which consists of more than 600,000 aligned Farsi–English sentences gathered from film subtitles. Experimental results show that TEP works better than Mizan as a training corpus for SMT.

3 TEP++

TEP++ is a refined version of TEP. TEP is a quite noisy corpus and it triggers several failures in the Farsi SMT pipeline. Besides the problem of noise because it was gathered from film subtitles, it is in SF. Accordingly it would be inappropriate to use an SMT system trained on SF data for the translation of FF. Unfortunately discrepancies between formal and informal Farsi structures are quite con-

[3] http://mabnasoft.com/english/parstrans/index.htm

[4] http://www.machinetranslation.ir/

[5] http://www.postchi.com/

[6] http://matrix.statmt.org/resources/pair?l1=fa&l2=en#pair

[7] However other Farsi MT engines like the Shiraz system (Amtrup et al., 2000) or that of Mohaghegh (2012) use their own in-house datasets. As we are not able to replicate them we do not include them in our comparisons.

[8] The project started in 1998 and the last version was released in 2010 (http://nl.ijs.si/ME)

[9] http://catalog.elra.info/product_info.php?products_id=1111

[10] https://translate.google.com/

[11] A research group supported by the Iran Supreme Council of Information to provide data resources for Farsi language and speech processing (http://www.dadegan.ir)

siderable. In what follows we show some of these cases and try to illustrate the main challenges with refinements to TEP.

In terms of orthography, Farsi is one of the hardest languages. It is written with the Perso-Arabic script. Unlike Arabic, some Persian words have inter-word zero-width non-joiner spaces (or semi-spaces) (Rasooli et al., 2013). Usually semi-spaces are incorrectly written as regular space character (U+0020 and U+200c are the Unicode for space and semi-space, respectively) that can easily change the meaning of the constituent and even the syntax of the whole sentence. As an example the right form of the word greedy is آستین‌دراز ≡ /āstin-derāz/[12] with a semi-space character (between n sound and d sound). If it is written with a space as in آستین دراز ≡ /āstin e derāz/, it means long sleeve, a completely different meaning which will mislead the SMT engine. Another problem is the presence of multiple writing forms for some characters. For the character ی ≡ /y/ all forms of ی, ي and ئ are common. This inconsistent writing style exists similarly for several other characters. The diacritic problem is another issue. Words can appear both with and without diacritics, like اخیراً or اخیرا ≡ /axiran/ (recently). Clearly, these problems should be resolved in preprocessing.

In addition, SF has its own specific problems, one being lexical variation. Some words occur in SF texts that do not have any counterpart in FF e.g. ایول ≡ /eyval/ (good job). Syntax in SF is also a problem. Farsi is an SOV language but in SF, versions of sentences with SVO and VOS order are both common. For example, علی نامه رو بخون ≡ /æli nāme ro bexoun/ (Ali, read the letter) is a standard SOV sentence, but both VOS (بخون نامه رو علی) and SVO (علی بخون نامه رو) forms are very normal; even in SF these look more natural than the SOV variant. In TEP++, we tried to correct the order and syntax of the sentences as much as possible which was very challenging. Not only the order but also the internal constituents of the sentences had to be changed. For example the verb بخون ≡ /bexoun/ (read) in SF is بخوان ≡ /bexān/ in FF or آمد ≡ /āmad/ (came) is the formal version of اومد ≡ /oumad/. These types of changes do not just happen to verbs. Other cases are even worse, e.g the right form of "for them" in FF is برای آنها ≡ /barāye ānhā/ which is written as براشون ≡ /barāšoun/ in SF (two FF words are packed in a single SF word). SF suffers from word ambiguity problem as well. A word like تو ≡ /to/ (you) which in formal texts is translated only into "you" (3^{rd}-singular person), can mean both "you" (1^{st} and 3^{rd}-singular person) and "inside" in SF.

Problems with SF are not limited to those discussed. However as a solution we cleaned the TEP data both automatically and manually. As a mandatory prerequisite of the refinement phase we applied knowledge of Farsi linguistics and developed a rule-based system for some of the cases. The rule-based system includes 17 general rules/templates. For the remainder a team of 20 native speaker of Farsi, manually edited the corpus. The result is TEP++ with 578,251 aligned sentences, with an average length of 7 for the English side and 9 for Farsi. It includes 4,963,693 English tokens (62,185 unique tokens) and 5,065,434 Farsi tokens (122,432 unique tokens). TEP++ covers 94% of the TEP and we neglected the remaining 6% because of the bad quality of the original TEP data.

4 Experiments

This section is divided into 3 subsections. The first part reports the BLEU scores for three main Farsi corpora, Mizan, TEP and TEP++. We also discuss the problems with Mizan in Section 4.1 and perform error analysis on the output translations, where it is used as the SMT training data. In the second part using TEP and TEP++ we carry out monolingual translation between SF and FF (SF2FF) and discuss some use-cases for this type of translation task. Finally in the last part we show how SF2FF boosts the SMT quality for Farsi and report our results on the IWSLT–2013 dataset providing a comparison with FBK's system.

4.1 Mizan, TEP and TEP++

To test the performance of our engines, they were trained using Mizan, TEP and TEP++. We used Moses (Koehn et al., 2007) with the default configuration for phrase-based translation. For the language modeling, SRILM (Stolcke and others,

[12]We used Wikipedia phonetic chart to show the spellings of Farsi words and - character to show the semi-space. http://en.wikipedia.org/wiki/Persian_phonology

2002) was used. The evaluation measure is BLEU (Papineni et al., 2002) and to tune the models, we applied MERT (Och, 2003). Table 1 summarizes our experimental results for the Mizan dataset. We evaluated with two types of language models, 3-gram (LM3) and 5-gram (LM5). Numbers for both before and after tuning are reported. For all experiments training, tuning and test sets were selected randomly from the main corpus. The size of the test set is 1,000 and the tuning set is 2000 sentences. Training set sizes are reported in tables. For all experiments BLEU scores for Google Translate are reported as a baseline.

	EN–FA		FA–EN	
	Before	After	Before	After
LM3	8.24	10.47	11.70	**13.35**
LM5	8.54	**10.53**	11.97	13.14
Google Translate	2.32		4.21	
Training set	1,016,758 parallel sentences			
Corpus	Mizan			

Table 1: Experimental Results for Mizan

From a system that is trained on almost 1M sentences, we might expect better performance. To try to gain some insight into the nature of the problem, we randomly selected 100 Farsi translations and compared them with the reference sentences. Based on the statistics of the error analysis for the subset of 100 translations, 3 main reasons of the failures present themselves:

1. In more than half of the cases (59%) the decoder does not find the correct translation of a given word. Wrong lexical choice is the most common problem for the translation.

2. Due to the rich morphology of Farsi 41% of the words are generally translated with slight errors in their forms. The problem, therefore, is wrong word formation on the target side (Farsi). To give an example translating verbs into the wrong tense or with the wrong affixes.

3. 33% of the constituents have reordering problems. Some times the translations are correct but are not in their right positions.

Such deficiencies do not only apply for Mizan; they are common in Farsi SMT (and SMT in general even), no matter what training data is. Studying the results of translation error analysis, Farzi and Faili (2015) confirm our findings.

Another issue which should be considered about the Farsi SMT evaluation is that Farsi is a free word-order language. When compiling the results of our experiments, we only had a single reference available against which the output from our various systems could be compared. Computing automatic evaluation scores when translating into a free word-order language in the single-reference scenario is somewhat arbitrary. We would expect a manual evaluation on a subset of sentences to confirm that the output translations are somewhat better than the automatic evaluation scores suggest.

Similar to Mizan we repeated the same experiments for the TEP and TEP++. Table 2 and Table 3 show the results of these related experiments. Two engines were trained using the TEP and TEP++ corpora. In order to provide a comparison between the two corpora used, tuning and test sets were selected in a way which mirror each other in both datasets, i.e. TEP sentences and their counterparts in TEP++.

	EN–FA		FA–EN	
	Before	After	Before	After
LM3	10.12	**12.14**	17.29	17.60
LM5	10.69	11.88	18.05	**18.57**
Google Translate	1.14		6.60	
Training set	609,085 parallel sentences			
Corpus	TEP			

Table 2: Experimental Results for TEP

	EN–FA		FA–EN	
	Before	After	Before	After
LM3	15.93	19.37	27.29	29.21
LM5	15.93	**19.60**	28.25	**29.74**
Google Translate	3.27		7.35	
Training set	575,251 parallel sentences			
Corpus	TEP++			

Table 3: Experimental Results for TEP++

As can be seen in the FA–EN direction we reached +11.17 (60%) improvement and in EN–FA direction the improvident is +7.76 (63.92%).[13]

[13]The best performance using TEP for FA–EN is 18.57, the best for TEP++ is 29.74 and the improvement of FA–EN direction is 60%

Another achievement is that even where using less data, the TEP++ engine performs better. TEP++ includes 94% of the TEP (§3) so even with about 33K fewer sentences pairs in the training set we obtained better results. The BLEU scores of TEP++ still are significantly better than the baseline (TEP) considering the results of paired bootstrap resampling (Koehn, 2004).[14]

This improvement is not odd and we were expecting such numbers. As it was studied in Rasooli at al. (2013) and Bertoldi et al. (2013) preprocessing and normalization have a considerable effect in Farsi SMT, as we explained in §3. Results from Google Translate is another confirmation to this issue. SF (the language of TEP) is an almost unknown language for Google Translate hence translation from/into this language will provide inappropriate results. Results are slightly better for TEP++ because the sentences are cleaner and more formal which are close to that of Google Translate. Finally it should be mentioned that Moses generally works much better than Google Translate for Farsi MT and the quality of Google Translate significantly decreases for long sentences.

4.2 SF2FF Results

Doing the refinements on TEP to produce TEP++ that as explained in §3, was very laborious. The by-product was a pair of corpora, one in SF and one in FF. We trained a phrase-based translation engine using these corpora in order to translate from SF into FF. The benefit of having such an engine is to produce the cleaned FF for free, as the TEP refinement was a costly process. Moreover, having a knowledge of Farsi linguistics was a prerequisite. This engine provides the same functionality with less cost and without applying linguistic knowledge. The trained engine works like a black box and carries out all the refinements. Similar to ours, Fancellu et al. (2014) have also worked on monolingual SMT between Brazilian and European Portuguese.

In the SF–FF direction we obtained 88.94 BLEU points and in the opposite direction systems works with BLEU score of 81.62. This process –more than an MT task– is a transformation in which words are converted into the normalized/correct forms and the order of constituents are changed in some cases. Accordingly BLEU numbers are high. SF2FF engine helps us to stablish a fully automated pipeline to make a large-scale bilingual Farsi corpus. Any type of data can be taken from the internet such as film subtitles or tweets that are usually noisy with informal writing conventions. SF2FF can normalize them. and the normalized version is good enough to be aligned with the English side (or any other language). To show the application of SF2FF and its performance, it was fed a test set from TEP (the same dataset we used in the TEP experiment). The data was normalized by SF2FF. Normalization helps to provide a more precise translation. The pipeline is illustrated in Figure 1. Selected sentences are in SF and the BLEU score for their translation by TEP is 18.57. If SF2FF translates them into FF they would be cleaner and much closer to the language of TEP++ and consequently the results of SMT would be better. Sentences in the two sets are counterpart of each other. The TEP++ engine obtains a BLEU score of 29.72 on the formal/clean version of the same sentences. If the noisy data is cleaned by SF2FF and is then translated by TEP++, the BLEU score rises to 25.36, i.e. SF2FF provides +6.79-point improvement. The BLEU score obtained the normalized data is significantly better and is 36% higher than that of the original data which demonstrates the efficiency of SF2FF.

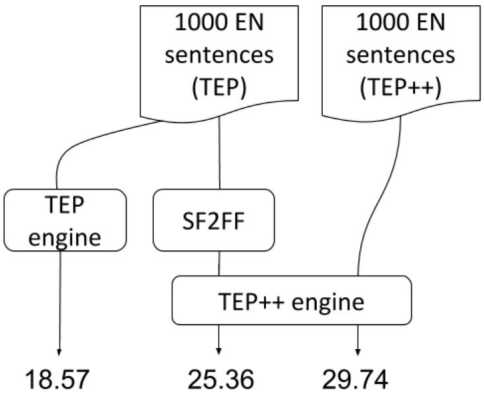

Figure 1: SF normalization by SF2FF

4.3 Comparison of SMT Performance

The only system that has been tested on a standard dataset and published is FBK's Farsi translation engine. It was reported in Bertoldi et al. (2013) and tested on the IWSLT–2013 dataset. The data has been made available by (Cettolo et al., 2012) and includes TED talk translations. In their paper,

[14]We used ARK research group codes for statistical significance testing for 1000 samples with 0.05 parameter http://www.ark.cs.cmu.edu/MT/

the FBK team explained that Farsi online data (including the IWSLT–2013 dataset) is very noisy and using requires some preprocessing, so they tried to normalize the data. Therefore, for the translation task, they used a normalized version of the IWSLT–2013 dataset along with an in-house corpus for language modeling. They also mentioned that using existing Farsi corpora such as TEP does not enhance translation quality. To compare our engines with FBK's system we firstly normalized the same dataset with SF2FF engine, and to make the language model we used the TEP++ corpus. The results for baseline,[15] FBK's system and ours (DCU) are shown in Table 4. For the FA–EN direction FBK obtained +2.0 points (16%) improvement in BLEU score, while for the same direction our improvement is +3.74 (29%). For the opposite direction we also outperform FBK, with a +1.10 difference in BLEU. The BLEU score for the EN–FA direction by DCU is 11.42, 2.29 points higher than the baseline (25%).

	Baseline	FBK	DCU
English-Farsi	9.13	10.32	**11.42**
Farsi-English	12.47	14.47	**16.21**

Table 4: Head-to-head comparison

5 Conclusion and Future Work

The contributions of this paper are threefold. First we developed a new corpus namely TEP++ and trained a translation engine. We showed that TEP++ works better than its predecessor TEP. Second we developed an engine to translate between FF and SF. SF2FF works like an intelligent preprocessor/normalizer and translates SF into FF that is a big credit for Farsi SMT. Finally we obtained better results in comparison to other reported results so far.

At the moment, in Farsi SMT data scarcity is the main challenge despite the fact that large volumes of textual data is available via the internet. Stored data on the internet for Farsi is in most cases are very noisy and also appears in SF forms. Our SF2FF engine can help to clean the internet data to generate reliable Farsi corpora. In the next step by normalizing existing Farsi corpora and aggregating them we will release a large-scale, reliable dataset for Farsi SMT. TEP++ also will be publicly available shortly. We also intended to carry out a human evaluation to investigate the correlation between the automatic score and manual findings.

[15] https://wit3.fbk.eu/score.php?release=2013-01

Acknowledgment

We would like to thank the three anonymous reviewers for their valuable comments. This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Dublin City University.

References

Amtrup, Jan Willers, Hamid Mansouri Rad, Karine Megerdoomian, and Rémi Zajac. 2000. *Persian–English machine translation: An overview of the Shiraz project*. Computing Research Laboratory, New Mexico State University, USA.

Bertoldi, Nicola, M Amin Farajian, Prashant Mathur, Nicholas Ruiz, and Marcello Federico. 2013. Fbks machine translation systems for the IWSLT 2013 evaluation campaign. In *Proceedings of the 10th International Workshop for Spoken Language Translation*. Heidelberg, Germany.

Cettolo, Mauro, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

Erjavec, Toma. 2010. Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Malta, European Language Resources Association (ELRA).

Farajian, Mohammad Amin. 2011. PEN: parallel English–Persian news corpus. In *Proceedings of the 2011th World Congress in Computer Science, Computer Engineering and Applied Computing*. Nevada, USA.

Farzi, Saeed and Heshaam Faili. 2015. A swarm-inspired re-ranker system for statistical machine translation. *Computer Speech & Language*, 29(1):45–62.

Federico, Fancellu, O'Brien Morgan, and Way Andy. 2014. Standard language variety conversion using smt. In *Proceedings of the Seventeenth Annual Conference of the European Association for Machine Translation (EAMT)*, pages 143–149, Dubrovnik, Croatia, May.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational*

Linguistics on Human Language Technology, pages 48–54. Edmonton, Canada.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Prague, Czech Republic.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pages 388–395, Barcelona, Spain.

Mohaghegh, Mahsa. 2012. *English–Persian phrase-based statistical machine translation: enhanced models, search and training, Massey University, Albany (Auckland), New Zealand*. Ph.D. thesis.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Sapporo, Japan.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Philadephia, Pennsylvania, USA.

Pilevar, Mohammad Taher, Heshaam Faili, and Abdol Hamid Pilevar. 2011. TEP: Tehran English–Persian parallel corpus. In *Computational Linguistics and Intelligent Text Processing*, pages 68–79. Springer.

Qasemizadeh, Behrang, Saeed Rahimi, and Behrooz Mahmoodi Bakhtiari. 2007. The first parallel multilingual corpus of persian: Toward a persian blark. In *The Second Workshop on Computational Approaches to Arabic Script-based Languages (CAASL-2)*. California, USA.

Rasooli, Mohammad Sadegh, Ahmed El Kholy, and Nizar Habash. 2013. Orthographic and morphological processing for Persian-to-English statistical machine translation. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 1047– 1051. Nagoya, Japan.

Stolcke, Andreas et al. 2002. SRILM an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904. Denver, Colorado.

Zajac, Rémi, Steve Helmreich, and Karine Megerdoomian. 2000. Black-box/glass-box evaluation in shiraz. In *Workshop on Machine Translation Evaluation at LREC-2000*. Athens, Greece.