Language Independent Connectivity Strength Features for Phrase Pivot Statistical Machine Translation

Ahmed El Kholy, Nizar Habash
Center for Computational Learning Systems, Columbia University
{akholy,habash}@ccls.columbia.edu

Gregor Leusch, Evgeny Matusov
Science Applications International Corporation
{gregor.leusch,evgeny.matusov}@saic.com

Hassan Sawaf
eBay Inc.
hsawaf@ebay.com

Abstract

An important challenge to statistical machine translation (SMT) is the lack of parallel data for many language pairs. One common solution is to pivot through a third language for which there exist parallel corpora with the source and target languages. Although pivoting is a robust technique, it introduces some low quality translations. In this paper, we present two language-independent features to improve the quality of phrase-pivot based SMT. The features, source connectivity strength and target connectivity strength reflect the quality of projected alignments between the source and target phrases in the pivot phrase table. We show positive results (0.6 BLEU points) on Persian-Arabic SMT as a case study.

1 Introduction

One of the main issues in statistical machine translation (SMT) is the scarcity of parallel data for many language pairs especially when the source and target languages are morphologically rich. A common SMT solution to the lack of parallel data is to pivot the translation through a third language (called pivot or bridge language) for which there exist abundant parallel corpora with the source and target languages. The literature covers many pivoting techniques. One of the best performing techniques, phrase pivoting (Utiyama and Isahara, 2007), builds an induced new phrase table between the source and target. One of the main issues of this technique is that the size of the newly created pivot phrase table is very large (Utiyama and Isahara, 2007). Moreover, many of the produced phrase pairs are of low quality which affects the translation choices during decoding and the overall translation quality. In this paper, we introduce language independent features to determine the quality of the pivot phrase pairs between source and target. We show positive results (0.6 BLEU points) on Persian-Arabic SMT.

Next, we briefly discuss some related work. We then review two common pivoting strategies and how we use them in Section 3. This is followed by our approach to using connectivity strength features in Section 4. We present our experimental results in Section 5.

2 Related Work

Many researchers have investigated the use of pivoting (or bridging) approaches to solve the data scarcity issue (Utiyama and Isahara, 2007; Wu and Wang, 2009; Khalilov et al., 2008; Bertoldi et al., 2008; Habash and Hu, 2009). The main idea is to introduce a pivot language, for which there exist large source-pivot and pivot-target bilingual corpora. Pivoting has been explored for closely related languages (Hajič et al., 2000) as well as unrelated languages (Koehn et al., 2009; Habash and Hu, 2009). Many different pivot strategies have been presented in the literature. The following three are perhaps the most common.

The first strategy is the sentence translation technique in which we first translate the source sentence to the pivot language, and then translate the pivot language sentence to the target language

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 412–418,
Sofia, Bulgaria, August 4-9 2013.

(Khalilov et al., 2008).

The second strategy is based on phrase pivoting (Utiyama and Isahara, 2007; Cohn and Lapata, 2007; Wu and Wang, 2009). In phrase pivoting, a new source-target phrase table (translation model) is induced from source-pivot and pivot-target phrase tables. Lexical weights and translation probabilities are computed from the two translation models.

The third strategy is to create a synthetic source-target corpus by translating the pivot side of source-pivot corpus to the target language using an existing pivot-target model (Bertoldi et al., 2008).

In this paper, we build on the phrase pivoting approach, which has been shown to be the best with comparable settings (Utiyama and Isahara, 2007). We extend phrase table scores with two other features that are language independent.

Since both Persian and Arabic are morphologically rich, we should mention that there has been a lot of work on translation to and from morphologically rich languages (Yeniterzi and Oflazer, 2010; Elming and Habash, 2009; El Kholy and Habash, 2010a; Habash and Sadat, 2006; Kathol and Zheng, 2008). Most of these efforts are focused on syntactic and morphological processing to improve the quality of translation.

To our knowledge, there hasn't been a lot of work on Persian and Arabic as a language pair. The only effort that we are aware of is based on improving the reordering models for Persian-Arabic SMT (Matusov and Köprü, 2010).

3 Pivoting Strategies

In this section, we review the two pivoting strategies that are our baselines. We also discuss how we overcome the large expansion of source-to-target phrase pairs in the process of creating a pivot phrase table.

3.1 Sentence Pivoting

In sentence pivoting, English is used as an interface between two separate phrase-based MT systems; Persian-English direct system and English-Arabic direct system. Given a Persian sentence, we first translate the Persian sentence from Persian to English, and then from English to Arabic.

3.2 Phrase Pivoting

In phrase pivoting (sometimes called triangulation or phrase table multiplication), we train a Persian-to-Arabic and an English-Arabic translation models, such as those used in the sentence pivoting technique. Based on these two models, we induce a new Persian-Arabic translation model.

Since we build our models on top of Moses phrase-based SMT (Koehn et al., 2007), we need to provide the same set of phrase translation probability distributions.[1] We follow Utiyama and Isahara (2007) in computing the probability distributions. The following are the set of equations used to compute the lexical probabilities (ϕ) and the phrase probabilities (p_w)

$$\phi(f|a) = \sum_e \phi(f|e)\phi(e|a)$$
$$\phi(a|f) = \sum_e \phi(a|e)\phi(e|f)$$
$$p_w(f|a) = \sum_e p_w(f|e)p_w(e|a)$$
$$p_w(a|f) = \sum_e p_w(a|e)p_w(e|f)$$

where f is the Persian source phrase. e is the English pivot phrase that is common in both Persian-English translation model and English-Arabic translation model. a is the Arabic target phrase.

We also build a Persian-Arabic reordering table using the same technique but we compute the reordering weights in a similar manner to Henriquez et al. (2010).

As discussed earlier, the induced Persian-Arabic phrase and reordering tables are very large. Table 1 shows the amount of parallel corpora used to train the Persian-English and the English-Arabic and the equivalent phrase table sizes compared to the induced Persian-Arabic phrase table.[2]

We introduce a basic filtering technique discussed next to address this issue and present some baseline experiments to test its performance in Section 5.3.

3.3 Filtering for Phrase Pivoting

The main idea of the filtering process is to select the top [n] English candidate phrases for each Persian phrase from the Persian-English phrase table and similarly select the top [n] Arabic target phrases for each English phrase from the English-Arabic phrase table and then perform the pivoting process described earlier to create a pivoted

[1] Four different phrase translation scores are computed in Moses' phrase tables: two lexical weighting scores and two phrase translation probabilities.

[2] The size of the induced phrase table size is computed but not created.

Translation Model	Training Corpora Size	Phrase Table	
		# Phrase Pairs	Size
Persian-English	≈4M words	96,04,103	1.1GB
English-Arabic	≈60M words	111,702,225	14GB
Pivot_Persian-Arabic	N/A	39,199,269,195	≈2.5TB

Table 1: Translation Models Phrase Table comparison in terms of number of line and sizes.

Persian-Arabic phrase table. To select the top candidates, we first rank all the candidates based on the log linear scores computed from the phrase translation probabilities and lexical weights multiplied by the optimized decoding weights then we pick the top [n] pairs.

We compare the different pivoting strategies and various filtering thresholds in Section 5.3.

4 Approach

One of the main challenges in phrase pivoting is the very large size of the induced phrase table. It becomes even more challenging if either the source or target language is morphologically rich. The number of translation candidates (fanout) increases due to ambiguity and richness (discussed in more details in Section 5.2) which in return increases the number of combinations between source and target phrases. Since the only criteria of matching between the source and target phrase is through a pivot phrase, many of the induced phrase pairs are of low quality. These phrase pairs unnecessarily increase the search space and hurt the overall quality of translation.

To solve this problem, we introduce two language-independent features which are added to the log linear space of features in order to determine the quality of the pivot phrase pairs. We call these features *connectivity strength features*.

Connectivity Strength Features provide two scores, Source Connectivity Strength (SCS) and Target Connectivity Strength (TCS). These two scores are similar to precision and recall metrics. They depend on the number of alignment links between words in the source phrase to words of the target phrase. SCS and TSC are defined in equations 1 and 2 where $\mathcal{S} = \{i : 1 \leq i \leq S\}$ is the set of source words in a given phrase pair in the pivot phrase table and $\mathcal{T} = \{j : 1 \leq j \leq T\}$ is the set of the equivalent target words. The word alignment between $\mathcal{S}$ and $\mathcal{T}$ is defined as $\mathcal{A} = \{(i, j) : i \in \mathcal{S} \text{ and } j \in \mathcal{T}\}$.

$$SCS = \frac{|\mathcal{A}|}{|\mathcal{S}|} \quad (1)$$

$$TCS = \frac{|\mathcal{A}|}{|\mathcal{T}|} \quad (2)$$

We get the alignment links by projecting the alignments of source-pivot to the pivot-target phrase pairs used in pivoting. If the source-target phrase pair are connected through more than one pivot phrase, we take the union of the alignments.

In contrast to the aggregated values represented in the lexical weights and the phrase probabilities, connectivity strength features provide additional information by counting the actual links between the source and target phrases. They provide an independent and direct approach to measure how good or bad a given phrase pair are connected.

Figure 1 and 2 are two examples (one good, one bad) Persian-Arabic phrase pairs in a pivot phrase table induced by pivoting through English.[3] In the first example, each Persian word is aligned to an Arabic word. The meaning is preserved in both phrases which is reflected in the SCS and TCS scores. In the second example, only one Persian word in aligned to one Arabic word in the equivalent phrase and the two phrases conveys two different meanings. The English phrase is not a good translation for either, which leads to this bad pairing. This is reflected in the SCS and TCS scores.

5 Experiments

In this section, we present a set of baseline experiments including a simple filtering technique to overcome the huge expansion of the pivot phrase table. Then we present our results in using connectivity strength features to improve Persian-Arabic pivot translation quality.

[3]We use the Habash-Soudi-Buckwalter Arabic transliteration (Habash et al., 2007) in the figures with extensions for Persian as suggested by Habash (2010).

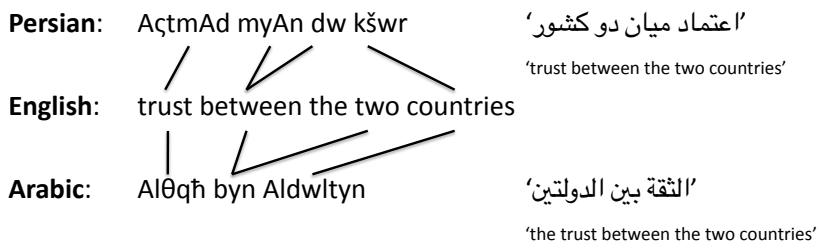

Figure 1: An example of strongly connected Persian-Arabic phrase pair through English. All Persian words are connected to one or more Arabic words. SCS=1.0 and TCS=1.0.

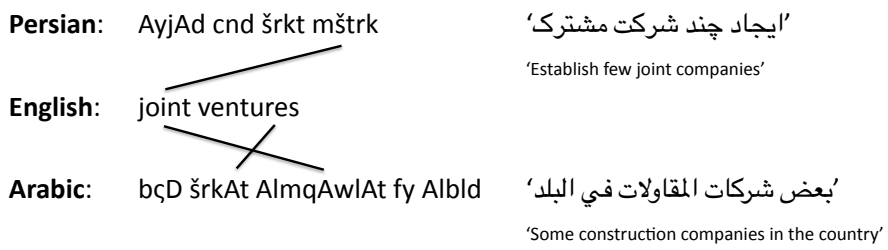

Figure 2: An example of weakly connected Persian-Arabic phrase pairs through English. Only one Persian word is connected to an Arabic word. SCS=0.25 and TCS=0.2.

5.1 Experimental Setup

In our pivoting experiments, we build two SMT models. One model to translate from Persian to English and another model to translate from English to Arabic. The English-Arabic parallel corpus is about 2.8M sentences (≈60M words) available from LDC[4] and GALE[5] constrained data. We use an in-house Persian-English parallel corpus of about 170K sentences and 4M words.

Word alignment is done using GIZA++ (Och and Ney, 2003). For Arabic language modeling, we use 200M words from the Arabic Gigaword Corpus (Graff, 2007) together with the Arabic side of our training data. We use 5-grams for all language models (LMs) implemented using the SRILM toolkit (Stolcke, 2002). For English language modeling, we use English Gigaword Corpus with 5-gram LM using the KenLM toolkit (Heafield, 2011).

All experiments are conducted using the Moses phrase-based SMT system (Koehn et al., 2007). We use MERT (Och, 2003) for decoding weight optimization. For Persian-English translation model, weights are optimized using a set 1000 sentences randomly sampled from the parallel corpus while the English-Arabic translation model weights are optimized using a set of 500 sentences from the 2004 NIST MT evaluation test set (MT04). The optimized weights are used for ranking and filtering (discussed in Section 3.3).

We use a maximum phrase length of size 8 across all models. We report results on an in-house Persian-Arabic evaluation set of 536 sentences with three references. We evaluate using BLEU-4 (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007).

5.2 Linguistic Preprocessing

In this section we present our motivation and choice for preprocessing Arabic, Persian, English data. Both Arabic and Persian are morphologically complex languages but they belong to two different language families. They both express richness and linguistic complexities in different ways.

One aspect of Arabic's complexity is its various attachable clitics and numerous morphological features (Habash, 2010). We follow El Kholy and Habash (2010a) and use the PATB tokenization scheme (Maamouri et al., 2004) in our

[4]LDC Catalog IDs: LDC2005E83, LDC2006E24, LDC2006E34, LDC2006E85, LDC2006E92, LDC2006G05, LDC2007E06, LDC2007E101, LDC2007E103, LDC2007E46, LDC2007E86, LDC2008E40, LDC2008E56, LDC2008G05, LDC2009E16, LDC2009G01.

[5]Global Autonomous Language Exploitation, or GALE, is a DARPA-funded research project.

experiments. We use MADA v3.1 (Habash and Rambow, 2005; Habash et al., 2009) to tokenize the Arabic text. We only evaluate on detokenized and orthographically correct (enriched) output following the work of El Kholy and Habash (2010b).

Persian on the other hand has a relatively simple nominal system. There is no case system and words do not inflect with gender except for a few animate Arabic loanwords. Unlike Arabic, Persian shows only two values for number, just singular and plural (no dual), which are usually marked by either the suffix ها+ *+hA* and sometimes ان+ *+An*, or one of the Arabic plural markers. Verbal morphology is very complex in Persian. Each verb has a past and present root and many verbs have attached prefix that is regarded part of the root. A verb in Persian inflects for 14 different tense, mood, aspect, person, number and voice combination values (Rasooli et al., 2013). We use Perstem (Jadidinejad et al., 2010) for segmenting Persian text.

English, our pivot language, is quite different from both Arabic and Persian. English is poor in morphology and barely inflects for number and tense, and for person in a limited context. English preprocessing simply includes down-casing, separating punctuation and splitting off "'s".

5.3 Baseline Evaluation

We compare the performance of sentence pivoting against phrase pivoting with different filtering thresholds. The results are presented in Table 2. In general, the phrase pivoting outperforms the sentence pivoting even when we use a small filtering threshold of size 100. Moreover, the higher the threshold the better the performance but with a diminishing gain.

Pivot Scheme	BLEU	METEOR
Sentence Pivoting	19.2	36.4
Phrase_Pivot_F100	19.4	37.4
Phrase_Pivot_F500	20.1	38.1
Phrase_Pivot_F1K	**20.5**	**38.6**

Table 2: Sentence pivoting versus phrase pivoting with different filtering thresholds (100/500/1000).

We use the best performing setup across the rest of the experiments.

5.4 Connectivity Strength Features Evaluation

In this experiment, we test the performance of adding the connectivity strength features (+*Conn*) to the best performing phrase pivoting model (*Phrase_Pivot_F1K*).

Model	BLEU	METEOR
Sentence Pivoting	19.2	36.4
Phrase_Pivot_F1K	20.5	38.6
Phrase_Pivot_F1K+Conn	**21.1**	**38.9**

Table 3: Connectivity strength features experiment result.

The results in Table 3 show that we get a nice improvement of ≈0.6/0.5 (BLEU/METEOR) points by adding the connectivity strength features. The differences in BLEU scores between this setup and all other systems are statistically significant above the 95% level. Statistical significance is computed using paired bootstrap resampling (Koehn, 2004).

6 Conclusion and Future Work

We presented an experiment showing the effect of using two language independent features, source connectivity score and target connectivity score, to improve the quality of pivot-based SMT. We showed that these features help improving the overall translation quality. In the future, we plan to explore other features, e.g., the number of the pivot phases used in connecting the source and target phrase pair and the similarity between these pivot phrases. We also plan to explore language specific features which could be extracted from some seed parallel data, e.g., syntactic and morphological compatibility of the source and target phrase pairs.

Acknowledgments

The work presented in this paper was possible thanks to a generous research grant from Science Applications International Corporation (SAIC). The last author (Sawaf) contributed to the effort while he was at SAIC. We would like to thank M. Sadegh Rasooli and Jon Dehdari for helpful discussions and insights into Persian. We also thank the anonymous reviewers for their insightful comments.

References

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. 2008. Phrase-based statistical machine translation with pivot languages. *Proceeding of IWSLT*, pages 143–149.

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 728.

Ahmed El Kholy and Nizar Habash. 2010a. Orthographic and Morphological Processing for English-Arabic Statistical Machine Translation. In *Proceedings of Traitement Automatique du Langage Naturel (TALN-10)*. Montréal, Canada.

Ahmed El Kholy and Nizar Habash. 2010b. Techniques for Arabic Morphological Detokenization and Orthographic Denormalization. In *Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.

Jakob Elming and Nizar Habash. 2009. Syntactic Reordering for English-Arabic Phrase-Based Machine Translation. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, pages 69–77, Athens, Greece, March.

David Graff. 2007. Arabic Gigaword 3, LDC Catalog No.: LDC2003T40. Linguistic Data Consortium, University of Pennsylvania.

Nizar Habash and Jun Hu. 2009. Improving Arabic-Chinese Statistical Machine Translation using English as Pivot Language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 173–181, Athens, Greece, March.

Nizar Habash and Owen Rambow. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573–580, Ann Arbor, Michigan.

Nizar Habash and Fatiha Sadat. 2006. Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 49–52, New York City, USA.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In Khalid Choukri and Bente Maegaard, editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*. The MEDAR Consortium, April.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Jan Hajič, Jan Hric, and Vladislav Kubon. 2000. Machine Translation of Very Close Languages. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP'2000)*, pages 7–12, Seattle.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, UK.

Carlos Henriquez, Rafael E. Banchs, and José B. Mariño. 2010. Learning reordering models for statistical machine translation with a pivot language.

Amir Hossein Jadidinejad, Fariborz Mahmoudi, and Jon Dehdari. 2010. Evaluation of PerStem: a simple and efficient stemming algorithm for Persian. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, pages 98–101.

Andreas Kathol and Jing Zheng. 2008. Strategies for building a Farsi-English smt system from limited resources. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH2008)*, pages 2731–2734, Brisbane, Australia.

M. Khalilov, Marta R. Costa-juss, Jos A. R. Fonollosa, Rafael E. Banchs, B. Chen, M. Zhang, A. Aw, H. Li, Jos B. Mario, Adolfo Hernndez, and Carlos A. Henrquez Q. 2008. The talp & i2r smt systems for iwslt 2008. In *International Workshop on Spoken Language Translation. IWSLT 2008, pg. 116–123.*

Philipp Koehn, Hieu Hoang, Alexandra Birch, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 machine translation systems for europe. *Proceedings of MT Summit XII*, pages 65–72.

Philipp Koehn. 2004. Statistical significance tests formachine translation evaluation. In *Proceedings of the Empirical Methods in Natural Language Processing Conference (EMNLP'04)*, Barcelona, Spain.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus.

In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.

Evgeny Matusov and Selçuk Köprü. 2010. Improving reordering in statistical machine translation from farsi. In *AMTA The Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado, USA.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–52.

Franz Josef Och. 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. Association for Computational Linguistics.

Kishore Papineni, 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Mohammad Sadegh Rasooli, Manouchehr Kouhestani, and Amirsaeid Moloodi. 2013. Development of a Persian syntactic dependency treebank. In *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, Atlanta, USA.

Andreas Stolcke. 2002. SRILM - an Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York, April. Association for Computational Linguistics.

Hua Wu and Haifeng Wang. 2009. Revisiting pivot language approach for machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 154–162, Suntec, Singapore, August. Association for Computational Linguistics.

Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntax-to-morphology mapping in factored phrase-based statistical machine translation from english to turkish. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 454–464, Uppsala, Sweden, July. Association for Computational Linguistics.