Graphic Language Model for Agglutinative Languages: Uyghur as Study Case

Miliwan Xuehelaiti[1,2,3] *, Kai Liu[2], Wenbin Jiang[2], and Tuergen Yibulayin[1]

1. Xinjiang University Information science and Technology institute
Urumqi,Xinjiang 830046,China
mihreban@126.com
2. Institute of Computing Technology, University of Chinese Academy of Sciences
Beijing 100190,China
{liukai,jiangwenbin}@ict.ac.cn
3. Urumqi Administration of Industry and Commerce
Urumqi,Xinjiang 830002,China

Abstract. This paper describes a novel, graphic language modeling strategy for morphologically rich agglutinative languages. Different from the linear structure in n-gram language models, graphic modeling organizes the morphemes in a sentence, including stems and affixes, as a directed graph. The graphic language model is verified in two typical application scenarios, morphological analysis and machine translation. We take Uyghur for example, and experiments show that the graphic language model achieves significant improvement in both morphological analysis and machine translation.

Keywords: graphic language model, agglutinative language, morphological analyzing, statistical machine translation

1 Introduction

Language model is one of the most important models in NLP, it describes probabilities of sentences in natural language. Many NLP tasks can be boiled down to the modeling of language model, such as transliteration, speech recognition, part of speech tagging and so on. One of most widely used language model is n-gram language model[3], which models words in sentences with local context environment in an linear way. The n-gram language model is simple and effective, and it have got excellent performance on Chinese, English and other languages with simple morphological form.

Agglutinative language is one kind of language which is widely used in North/South Korea, Japan, Mongolian, Turkey and other countries in Middle East Asia and other areas. Agglutinative languages differ from languages with simple morphological form (such as English and Chinese) in their sentence and

* This work is supported by the National Science Foundation of China (Grant No. 61262060 61262060), Key Project of National Natural Science Fund [61032008] and National Social Science Fund Key Projects [10AYY006].

word-formation[2], in which the composition of each word in agglutinative language follows different word-building rules according to simple observation: each word of agglutinative language is composed by a word-stem and any number of affixes, in where constrained relations exist between stem and affixes; and similar relations exist in stems of different words. The former rule lead to the data sparseness of word of agglutinative language, the latter rule makes it hard to seize the relation between stems, for there maybe some affixes between stems in different words.

According to the observations above, sentence of agglutinative language with those relations can not be simply modeled as linear sequence. As a matter of fact, traditional n-gram language language model which models sentences as linear sequence can not obtain idea results on agglutinative languages. In this paper, we propose a novel graphic language model which can depict those relations more deeply. More specifically, our graphic language model models the generative relations between stem and affixes in a word and the relations between stems in different words, which relations can hardly be modeled by traditional linear language models.

In order to test the novel graphic language model, two language model needed natural language processing tasks (morphological analyzing and statistical machine translation(SMT)) are adopted to verify our graphic language model. In the experiments, both tasks show that graphic language model gets significant improvements compares to n-gram language model. In morphological analyzing, the accuracy gains 0.8% improvements due to the new style language model, while in SMT it gains more than 1.1 BLEU improvement. Furthermore, the graphic language model is simple, and the complexity of it is approximate to the n-gram language model.

The rest of paper is organized as follows: Section 2 describes the characteristics of agglutinative languages; Traditional linear language model is described in Section 3; We propose our graphic language model for agglutinative language in Section 4; And the methods of utilizing proposed graphic language model in two NLP tasks are shown in Section 5; Finally, we present the experiments of the two NLP tasks with graphic language model on Uyghur in Section 6 and conclude in Section 8.

2 Agglutinative Language

Agglutinative language is a kind of language that its words are made up of distinct morphemes by a linear sequence way, and each component of meaning is represented by its own morpheme. Agglutinative languages have many characteristic, more specifically, we take Uyghur as our study case. Uyghur is one of typical agglutinative languages, it is a Turkic language which is widely used in Western China by Uyghur people, and it shares some characteristics with other agglutinative languages:

– Each word jointed with different affixes will show different meanings. Take a word in Uyghur as example, the word "xizmet"(job) will show differen-

Word	Stemming	Meaning
Ölchem	*Ölchem*	standard
Ölchemlesh	*Ölchem+lesh*	standardization
Ölchemleshtür	*Ölchem+lesh+ür*	standardize (it)
Ölchemleshtürel	*Ölchem+lesh+ür+el*	can standardize (it)
Ölchemleshtürelme	*Ölchem+lesh+ür+el+me*	can not standardize (it)
Ölchemleshtürelme	*Ölchem+lesh+ür+el+me+m*	can not standardize (it)?
Ölchemleshtürelmesiler	*Ölchem+lesh+ür+el+me+m+siler*	can't you standardize (it)?

Table 1. An example of agglutinative language's word with multiple morphemes. A word with different morphemes will show different meanings and even be a short sentence.

t meanings when different morphemes followed with it: "xizmettin"(from work), "xizmetde"(with work) and so on.
- Each word can be jointed with multiple morphemes, and such a word can even be a short sentence. As it is shown in Table 1, the same word "*Ölchem*" (standard) jointed with different morphemes convey different meanings. And much important information, such as, meanings of content words are conveyed by those morphemes.

In addition, morphemes of Uyghur fall into two categories: stem and affix. The first morpheme of each word is the stem of the word in Uyghur, and each word should have one and only one stem, which conveys the main semantic meaning of the word. As the example above, "xizmet"(job) and "*Ölchem*"(standard) are stems. And all morphemes after stems are affixes, which convey minor semantic meanings or grammar information. Furthermore, all stems in a sentence form the skeleton of the sentence.

3 Linear Language Model

Language model describes a word sequence w_j^i $(w_i, w_{i+1}, ..., w_{j-1}, w_j)$ by assigning the probability $P(w_j^i)$ to the sequence by means of certain probability distribution. And language model is widely used in many NLP applications, such as speech recognizing, morphological analyzing, machine translation and so on.

Most language models regard the word sequence as a linear structure, and calculate the probability in a linear way. One of the most typical models is n-gram language model, which assigns a given word's probability by means of its previous contiguous words. In n-gram language model, the probability of the sequence w_j^i is assigned approximated as:

$$P(w_j^i) = \prod_k P(w_k|w_{k-1}^i) \approx \prod_k P(w_k|w_{k-1}^{k-n+1}) \tag{1}$$

	Uyghur	**Chinese**
total word	69563	43285
total freq	1263861	1161801
avg freq	18.17	26.84

Table 2. The lexical statistical data for Uyghur and Chinese. Total word: the total number of words existed in the corpus; total freq: the sum of all words' frequencies; avg freq: the average frequency of each word.

where n is the n-gram size of the language model. And $P(w_k|w_{k-1}^{k-n+1})$ can be calculate from its frequency in corpus:

$$P(w_k|w_{k-1}^{k-n+1}) = \frac{\#(w_k^{k-n+1})}{\#(w_{k-1}^{k-n+1})} \tag{2}$$

where $\#(\cdot)$ means the total count of the n-gram in the corpus. And in practical terms, the probability above needs some kind of smoothing, such as "add-one", Good-Turing, Kneser-Ney smoothing and so on[4].

3.1 Linear modeling for Agglutinative Language

Because of the characteristic of agglutinative language, common linear modeling for agglutinative language will encounter some problems:

- Data sparseness. For each word in agglutinative language can be made of several morphemes, and the number of all probable words are astronomical, which can make serious data sparseness problem. Take Uyghur and Chinese as example, we count the lexical frequencies of both languages in parallel corpus (Table 2), from where we can see less frequency per word in Uyghur, in other words, more data spareness.
- Ignoring the relations in/between words. Traditional linear language model ignore to model relations between stems and relations inside the word. And a remedial measure for it is to model morphemes as basic units instead of words in n-gram model. This measure can describe stem-affix relations, but it still have weaker ability to describe the relations between stems.

4 Graphic Language Model

Since there are several drawbacks of linear modeling agglutinative language, we propose a morpheme based directed graphic language model. And the directed graphic structure can better describe the characteristic of agglutinative language.

As it is shown in Figure 1, we model a agglutinative language sentence as directed graphic with morphemes as basic elements, where all stems are connected linearly in left-right order and all affixes are connected to previous affixes or stems. And it can be divided into two linear parts:

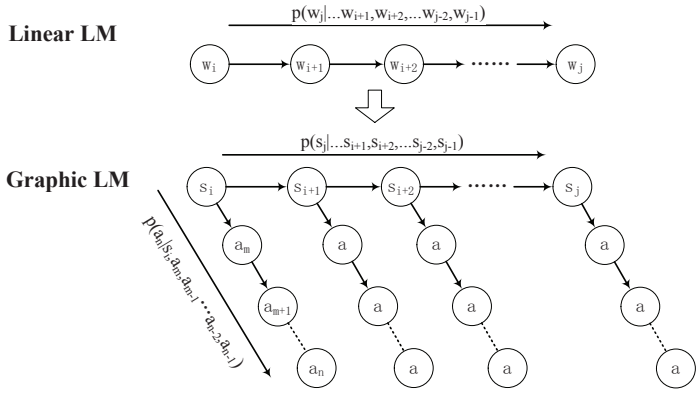

Fig. 1. The structures of linear language model (Linear LM) and graphic language model (Graphic LM). Compared to linear language model, graphic language model describe the detail relation of stems and affixes inside/outside the words. And as it is shown in the figure, it can be separated into two parts of sub-models(stem-stem and stem-affix).

- stem-stem: this part assigns the probability to all stems, similar to n-gram model, the probability can be calculated approximately by assuming the probability of observing stem can be calculated in condition of preceding n stems:

$$P(s_i^1) = \prod_i P(s_i|s_{i-1}^1) \approx \prod_i P(s_i|s_{i-1}^{i-n+1}) \tag{3}$$

  where s_i denotes the stem of i^{th} word.
- stem-affix: this part calculates all affixes probabilities by means of preceding stem and affixes in the same word:

$$P(a) = \prod_i \prod_j P(a_{i,j}|s_i, a_{i,j-1}^{i,1}) \tag{4}$$

  where $a_{i,j}$ denotes the j^{th} affix in i^{th} word in the sentence.

For the whole model, we combine both parts above together, and calculate sentences' probabilities as follows:

$$P(w_n^1) = \prod_i (P(s_i|s_{i-1}^{i-n+1}) \prod_j P(a_{i,j}|s_i, a_{i,j-1}^{i,1})) \tag{5}$$

where stems s and affixes a are obtained from morphological analyzing results of the sentence w_n^1. The training method of both parts of this model can refer to n-gram language model. In this way, we design a morpheme based directed graphic language model, which is supposed to be a better language model for agglutinative language.

Algorithm 1 Estimating Probability of Sequence

1: Stemmed Seq←MA(Seq)
2: Stem-Stem Seq←RemoveAffix(Stemmed Seq)
3: Stem-Affix Seq List←Split(Stemmed Seq)
4: Stem-Stem Prob←LM(Stem-Stem Seq, Stem-Stem Model)
5: **for each** Stem-Affix Seq ∈ Stem-Affix Seq List **do**
6: Stem-Affix Prob←LM(Stem-Affix Seq, Stem-Affix Model)
7: **end for**
8: Graphic Prob←Multiply(Stem-Stem Prob, Stem-Affix Prob)

Training Firstly, we morphological analyze the training corpus into stemmed one. Secondly, according to the stemmed corpus, we obtain stem-stem and stem-affix corpora by removing affixes and splitting sentences by words respectively. Then, those corpora are utilized to train linear language model for stem-stem and stem-affix respectively.

Estimating Probability Algorithm 1 outlines the estimation procedure in its entirety. In line 1, we analyze the input sequence by morphological analyzing procedure $MA(\cdot)$ and obtain the stemmed corpus. The stemmed corpus is utilized to obtain stem-stem sequences by removing all affixes in the corpus in line 2. Correspondingly, in line 3 sentences are split according to words and then organized into stem-affix sequences. From line 4-7, we calculate the both parts' probabilities with corresponding sequences and sub models through procedure $LM(\cdot,\cdot)$. And the final score of the model is combined by scores from sub-models in line 8.

5 Applications

5.1 Morphological Analyzing

Morphological analyzing is one of the most important NLP tasks in agglutinative language[1]. The quality of morphological analyzing will affect other NLP tasks, which are based on morphological analyzing. There are three sub-task in morphological analyzing, including stemming, restoring the changed letter and POS tagging, in which we select the first sub-task stemming as the application to verify our graphic language model. Stemming is similar to segmentation, it splits each word into morphemes (Figure 2), including a stem and several following affixes. According to the characteristic of agglutinative language, stemming needs the contexts of inside or outside the word, where language model is available and important.

Formally, we define a word sequence as w_j^i, which means it is a word sequence with words from position i to j in sentence. And each word can be segmented into several morphemes m, which contain a single stem s and several following affixes a. In this task, we try to find the most probable morphological segmentation $m_{n'}^1$ of sentence w_n^1.

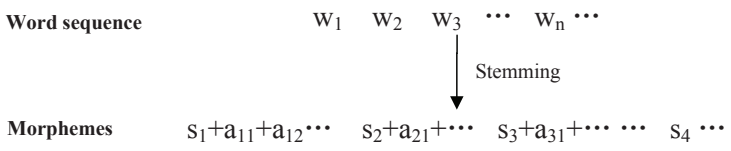

Fig. 2. Stemming word sequence into morphemes, where the first morpheme in each word is stem and others are affixes.

With linear modeling, n-gram language model selects the morpheme sequence with the maximum language model probability:

$$\ell(w_n^1) = \arg\max_{m_{n'}^1} \prod_i P(m_i|m_{i-1}^{i-n+1}) \tag{6}$$

where m denotes a morpheme, and $m_{n'}^1$ is the selected morpheme sequence of the sentence w_n^1. And the $P(m_i|m_{i-1}^{i-n+1})$ means n-gram language model's probability of morpheme m_i with context m_{i-1}^{i-n+1}.

With graphic modeling, correspondingly, we try to find the morpheme sequence with the maximum model probability with stems and affixes:

$$\ell(w_n^1) = \arg\max_{s,a} \prod_i P(s_i|s_{i-1}^0) \prod_{i,j} P(a_{i,j}|s_i, a_{i,j-1}^{i,0}) \tag{7}$$

where s_i denotes the stem of the i^{th} word, and $a_{i,j}$ denotes the j^{th} affix of i^{th} word. The first term of Formula 7 is the stem-stem part of our graphic model and the second term is the stem-affix part.

5.2 Machine Translation

Machine translation is one of the hardest problems in NLP. The performance of statistical machine translation is highly depended on the quality of the language model (cite), and it is a good task to verify the quality of language model. In this paper, we try to verify the effectiveness of graphic language model with agglutinative language as target side.

Due to the characteristics of agglutinative language, the application will be performed on two different SMT system with different granularities:

Word Based The SMT model is trained with words in agglutinative language side as the basic translation unit, and this kind of SMT system has several characteristics:

- It has large-grained translation unit, which may suffer from data sparseness problem.
- Shorter sequences of agglutinative language will be translated while we use word as basic translation unit.
- Word based translation system is free to recombination of morphemes into words.

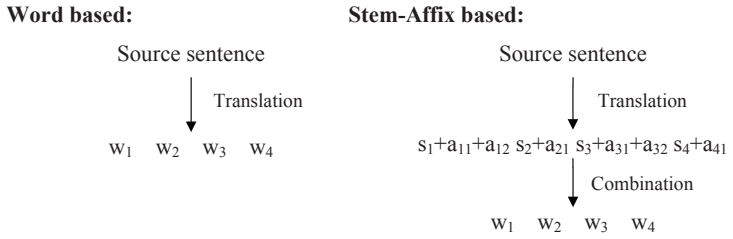

Fig. 3. Translation from one language to an agglutinative language with different basic units. Left panel shows the translation from source language directly to target words, while the right panel shows the translation procedure that translate source language firstly into morphemes and then combine them into final words.

Morpheme Based Correspondingly, the stemmed sentences are used to train the SMT model and stems and affixes are the basic translation unit here:

- Smaller-grained translation unit means less data sparseness problem.
- Longer sequences have to be translated while stems and affixes are chosen, which means higher translation complexity.
- Morpheme based translation system have to recombine stems and affixes into agglutinative words.

6 Experiments

In this section, we verify our graphic language model through two applications. And there are three different types of language model will be utilized in the experiments:

- Word linear LM: the n-gram language model based on words, and it will be utilized in the application of SMT.
- Morphemes linear LM: morpheme based n-gram language model which will be utilized in both applications.
- Graphic LM: our graphic language model for agglutinative language, and the order (n) of linear part is equal to corresponding comparison LM in experiments as default.

6.1 Morphological Analyzing

DataSet We make use of an annotated corpus Mega-words Corpus of Morphological Analysis of Uyghur, which is manually annotated by Xinjiang multilingual key laboratory. And it contains about 67 thousands sentences, from which we select 5% as our testing set.

Model	P%	R%	F%
Morpheme linear LM	87.5	87.4	87.5
Graphic LM	88.0	88.6	88.3

Table 3. Experiment results on morphological analyzing with different language model.

Training and Evaluation We train a 5-gram language model on morphemes and the linear part of our graphic model by SRI Language Modeling Toolkit[9] with Kneser-Ney smoothing. And we simply evaluate the stemming result by precision and recall of the morphemes.

Results As the results shown in Table 3, our graphic language model shows advantage on morphological analyzing compared to morpheme linear language model, where precision obtain an improvement of 0.5% and more than 1% improvement on recall.

6.2 Machine Translation

In this section, we compare our graphic language model with linear n-gram language model in SMT task. In SMT task, two different granularities are employed to verify the effectiveness of our graphic language model. One experiment is performed on word, while the other is performed on the results of morphological analyzing (stems and affixes).

DataSets For bilingual training data, we select Chinese-Uyghur corpus with 120 thousand parallel sentence pairs, which includes fifty thousand sentence pairs from corpus provided by CWMT 2011 evaluation task[5]. We obtain morphological result of the corpus by performing Uyghur morphological analyzer[1] on the corpus. The parallel corpus's word alignments are obtained by running GIZA++[6] on the corpus in both directions and applying "grow-diag-and" refinement.

Training and Evaluation We use the development set provided by CWMT 2011[2] evaluation task as our development set, and we organize 1000 sentence pairs as our own test set. The quality of translation is evaluated by the NIST BLEU-4 metric[7]. We make use of the standard MERT as the tuning algorithm to tune our cascaded translation model's parameters on development set.

[1] Developed by Institute of Computing Technology(ICT), Chinese Academy of Sciences(CAS)

[2] http://www.chineseldc.org/resource_info.php?rid=156

Granularity	Language Model	BLEU%
Word based	word5	51.19
	word5+morpheme5	52.28
	word5+stem3	53.01 (+0.73)
	word5+stem5	53.18 (+0.90)
	stem3+affix3	53.18 (+0.90)
	stem5+affix5	53.44 (+1.16)
Morpheme based	morpheme5	54.26
	stem3+affix3	54.91 (+0.65)
	stem5+affix5	55.26 (+1.00)

Table 4. Experiment results on test set. We test translation model with different language model respectively. word: word based n-gram language model; stem: stem part of our graphic language model; affix: affix part of our graphic language model; morpheme: the morpheme based linear language model. And stem+affix is our whole graphic language model. The followed number denotes the order of the language model, for example, word5 means a 5-gram word based language model and stem3 means we train the stem part with order 3.

Baselines and Our model We apply SRI Language Modeling Toolkit to train language models with modified Kneser-Ney smoothing on Uyghur side of the training corpus. The open source SMT decoder Moses[8] is selected as our baseline, which contains implementation of hierarchical phase model (Moses-chart). Correspondingly, our model is based on the same decoding system Moses, and train our graphic language model on the same corpus (training corpus).

Results The experiment result is shown in Table 4, which line 2-7 show the results of word based SMT model with different language model respectively. And line 9-10 give the results of morpheme based SMT model with both linear language model and our graphic language model.

As the results shown, our graphic language model is significant better than those linear modeling language model. And both parts of our graphic language model (stem-stem, stem-affix) show their effectiveness on experiment, while the whole model shows better performance. Meanwhile, those improvements prove that it is reasonable to model agglutinative language in stem-stem and stem-affix style, and this style of structure can describe some kinds characteristic of agglutinative language.

7 Related Work

Language Model In addition to n-gram language model, there are much work is devoted into language model. Some kind of structured language model aims at modeling the structures of language and overcoming the locality problem[10] and neural network is employed to improve the work[11]. But so far, there is not any work on the structure of agglutinative language.

Morphological Analyzing There are a lot of supervised work on morphological analyzing for each language respectively: Japanese[12], Arabic[13], and so on. Correspondingly, unsupervised ones (e.g.[14]) are also available. And morphological analyzing is proved to be an important task for other NLP task (e.g. SMT [16–18]).

Machine Translation So far, most studies of agglutinative related machine translation are base on agglutinative to non-agglutinative translation, such as, for Turkish[15–17], Korean[18, 19] and others[20]. And there is also work on alignment between agglutinative language and other languages in translation purpose[21, 22]. While there is less work on translation of non-agglutinative language to agglutinative language[23].

8 Conclusion and Future Work

In this paper, we model agglutinative language with graphic structure on the basis of the characteristics of agglutinative language by observations. The novel language model can better describe the agglutinative language and remit data sparseness, where evidences are provided by the experiments of different NLP tasks. The experiment results show significant improvements on morphological analyzing and SMT tasks with 0.8 F-score improvements and 1.1 BLEU improvements.

For future work, we will investigate other kinds of structures that can better model the agglutinative language (e.g. indirected graph or all connected graph) and involve more feature of agglutinative language into our model, or directly model the language model as discriminative model.

References

1. Arianna Bisazza and Marcello Federico. 2009. Morphological pre-processing for Turkish to English statistical machine translation. In Proceedings of IWSLT, pages 129C135.
2. K. Oflazer, Two-level description of Turkish morphology, in Literary and Linguistic Computing, 1994, vol. 9, no. 2, pp. 137C148.
3. Katz, Slava, Estimation of probabilities from sparse data for the language model component of a speech recognizer, Acoustics, Speech and Signal Processing, IEEE Transactions on, 1987, volume 35, 3, pages 400-401
4. Chen, Stanley F and Goodman, Joshua, An empirical study of smoothing techniques for language modeling, Proceedings of the 34th annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, 1996, pages 310-318
5. Hongmei Zhao, Yajuan L*ü*, Guosheng Ben, Yun Huang, Qun Liu, The evaluation report of CWMT2011, CWMT 2011, 2011, pages 261-180
6. F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. Computational linguistics, 29(1):19C51.

7. K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 311C318. Association for Computational Linguistics.
8. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In Annual meeting-association for computational linguistics, volume 45, page 2.
9. A. Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In Proceedings of the international conference on spoken language processing, volume 2, pages 901C904.
10. Chelba, Ciprian and Jelinek, Frederick, Structured language modeling, Computer Speech & Language, 2000, volume 14, 4, pages 283-332
11. Emami, Ahmad and Jelinek, Frederick, A neural syntactic language model, Machine learning,2005, volume 60, 1-3, pages 195-227, Springer
12. Nagata, Masaaki, A stochastic Japanese morphological analyzer using a forward-DP backward-A* N-best search algorithm, Proceedings of the 15th conference on Computational linguistics-Volume 1, pages 201–207, 1994, Association for Computational Linguistics
13. Buckwalter, Tim, Buckwalter {Arabic} Morphological Analyzer Version 1.0, 2002
14. Creutz, Mathias and Lagus, Krista, Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0, 2005, Helsinki University of Technology
15. Arianna Bisazza and Marcello Federico. 2009. Morphological pre-processing for Turkish to English statistical machine translation. In Proceedings of IWSLT, pages 129C135.
16. Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In Proceedings of HLT-EMNLP, pages 676C683.
17. Coskun Mermer and Murat Saraclar. 2011. Unsupervised Turkish morphological segmentation for statistical machine translation. In Workshop of MT and Morphologically-rich Languages.
18. Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In Proceedings of HLT-NAACL, Short Papers, pages 57C60.
19. Minh-Thang Luong, Preslav Nakov, and Min-Yen Kan. 2010. A hybrid morpheme-word representation for machine translation of morphologically rich languages. In Proceedings of EMNLP, pages 148C157.
20. Sami Virpioja, Jaakko J. Vayrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In Proceedings of MT SUMMIT, pages 491C498.
21. Minh-Thang Luong and Min-Yen Kan. 2010. Enhancing morphological alignment for translating highly inflected languages. In Proceedings of COLING, pages 743C751.
22. Zhiyang Wang, Yajuan Lu, and Qun Liu. 2011. Multi-granularity word alignment and decoding for agglutinative language translation. In Proceedings of MT SUMMIT, pages 360C367.
23. Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntaxto-morphology mapping in factored phrase-based statistical machine translation from English to Turkish. In Proceedings of ACL, pages 454C464.