LIG approach for IWSLT09 : Using Multiple Morphological Segmenters for Spoken Language Translation of Arabic

Fethi Bougares [#1], *Laurent Besacier* [#2], *Hervé Blanchon* [#3]

[#] LIG, University of Grenoble

[1] fethi.bougares@imag.fr

[2] laurent.besacier@imag.fr

[3] herve.blanchon@imag.fr

***Abstract*— This paper describes the LIG experiments in the context of IWSLT09 evaluation (Arabic to English Statistical Machine Translation task). Arabic is a morphologically rich language, and recent experimentations in our laboratory have shown that the performance of Arabic to English SMT systems varies greatly according to the Arabic morphological segmenters applied. Based on this observation, we propose to use simultaneously multiple segmentations for machine translation of Arabic. The core idea is to keep the ambiguity of the Arabic segmentation in the system input (using confusion networks or lattices). Then, we hope that the best segmentation will be chosen during MT decoding. The mathematics of this multiple segmentation approach are given. Practical implementations in the case of verbatim text translation as well as speech translation (outside of the scope of IWSLT09 this year) are proposed. Experiments conducted in the framework of IWSLT evaluation campaign show the potential of the multiple segmentation approach. The last part of this paper explains in detail the different systems submitted by LIG at IWSLT09 and the results obtained.**

1. Introduction

Spoken language translation of Arabic language has been widely studied recently in different projects (DARPA TRANSTAC, GALE) or evaluation campaigns (IWSLT[1], NIST[2]). Most of the time, the rich morphology of Arabic language is seen as a problem that must be addressed, especially when dealing with sparse data. It has been shown that pre-processing Arabic data using a morphological segmenter is useful to improve machine translation results [1] [2] or automatic speech recognition performances [3]. If such a strategy is applied, the choice of the Arabic segmenter is very important since the Arabic segmentation heavily influences the translation quality: segmentation affects the translation models (alignments, phrase table) as well as the translation input.

In a recent work [4] we conducted an in depth study of the influence of two Arabic segmenters on the translation quality of a phrase-based system using the *moses*[3] decoder. Examples of Arabic segmentations and associated translations are given in *table 1* where correct segmentations and correct translations (both evaluated by a human expert) are in bold. While the correct segmentation may lead to the correct translation (cases 1, 2 and 7), we also observed some sentences for which none of the proposed segmentations is correct (cases 3 and 4). In those cases, the translation output might still be correct. One reason may be that an incorrect segmentation can remain consistent with the segmentation applied on the training data (bad segmentation on the training data will probably lead to bad alignments but these errors may be somehow recovered during the phrase-table construction). Finally, we also observe cases (5 and 6) where a correct segmentation does not necessarily lead to the best translation output.

Table 1
Qualitative comparison of two Arabic segmentation methods (Buckwalter versus ASVM) for SMT. Correct segmentations and translations (human expertise) are bold-faced.

	Seg. Buckwalter	Seg. ASVM
1	ما مقاس ك **what size do you wear**	ما مقاسك what your size
2	س ي ستغرق ذلك حوالى ثلاثون دقيقة it will take that thirty minutes	سيستغرق ذلك حوالى ثلاثون دقيقة **it will take about thirty minutes**
3	أيمكنني استعمال هاتف ك **can i use your telephone**	أيمكنني استعمال هاتف ك can i your phone
4	عند ي حمى منذ ال بارح ة i have a fever since	عندي حمى منذ البارحة **i have a fever since yesterday**
5	هل ب إمكان ي ال تحدث إلى ال سيد كارتر can i talk to mr	هل ب إمكان ي التحدث إلى السيد كارتر **May i speak to mr carter**
6	صدم ت ني سيار ة **i was hit by a car**	صدمت ني سيارة Was hit by a car
7	عند ي ألم في ال أذن **i have a pain in my ear**	عند ي ألم في الأذن i have a pain in turn

Based on this analysis, we believe that using simultaneously multiple segmentations is a promising approach to improve machine translation of Arabic; this is the goal of the work described in this paper. A basic approach to implement this proposal would have been to build different MT systems using different segmentations of the Arabic training data and to combine their translations outputs. However, we think that it might be more interesting to leave the ambiguity of the Arabic segmentation at the input of the system (using a graph representation for instance). Then, the best segmentation should be chosen during the decoding step. We will describe this latter approach and discuss :

-the mathematics of this multiple segmentation approach (section 2),

[1] See for instance http://mastarpj.nict.go.jp/IWSLT2009/

[2] See http://www.itl.nist.gov/iad/mig//tests/mt/

[3] http://www.statmt.org/moses/

-a practical implementation in the case of verbatim text translation ; confusion networks (CN) are used to represent the ambiguity of the Arabic segmentation at the input of the MT system (section 3),

-problems and solutions to apply the multiple segmentation approach to spoken language translation using ASR lattices (section 4),

-experiments to validate the approach in the framework of IWSLT evaluation campaigns (section 5),

The last part of this paper (section 6) explains in detail the different systems submitted by LIG at IWSLT09 and the results obtained.

2. Formalisation of the multiple segmentation approach

Given *f* the Arabic sentence to be translated and a_1^k a particular morphological segmentation of *f*. The search of the best English translation *e* of *f* can be written as follows:

$$e^* = \arg\max_e P(e/f) \quad (1)$$

$$e^* = \arg\max_e \sum_{a_1^k} P(e, a_1^k / f) \quad (2)$$

$$e^* = \arg\max_e \sum_{a_1^k} P(a_1^k / f).P(e/f, a_1^k) \quad (3)$$

$$e^* \approx \arg\max_e \left\{ \max_{a_1^k} P(a_1^k / f).P(e/a_1^k) \right\} \quad (4)$$

From equation (3) to (4) we assume that the translation model is trained using an Arabic text segmented into morphemes (so *f* is removed) and the sum is approximated by a max function.

Then, the final equation, taking into account the multiple segmentation approach, is the following (after applying also Bayes rule and removing the denominator because of the max operator):

$$e^* \approx \arg\max_e \left\{ \max_{a_1^k} P(a_1^k / f).P(a_1^k / e).P(e) \right\} \quad (5)$$

where

$P(e)$ is the target language model,

$P(a_1^k / f)$ is the "segmentation" model,

$P(a_1^k / e)$ is the translation model trained for a given segmentation.

3. Multiple segmentation for verbatim translation

We are using the *moses* open source decoder which allows exploiting confusion networks (CN) as an interface data structure between speech recognition and machine translation [5]. CN decoding allows to represent a huge number of transcription hypotheses while leading to efficient search algorithms for statistical machine translation.

We also decided to use confusion networks (CN) to represent the ambiguity at the segmentation level[4]. *Figure 1* shows an example of a confusion network built for a sentence *f* segmented using 2 different morphological segmenters. The transitions correspond to different segmentation options. Probabilities can be associated to each transition but in this work, the different segmentation options are considered as equiprobable.

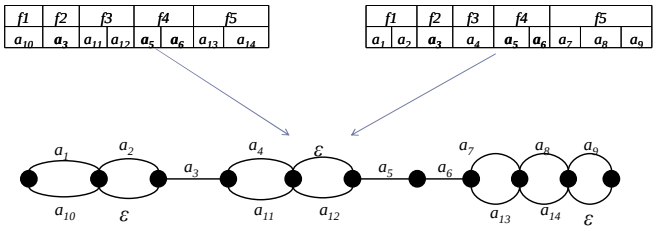

Fig. 1. Example of confusion network built for a sentence *f* segmented using 2 different morphological segmenters.

Then, the CN is decoded according to equation 5 using one target language model and several (two in this work) log-linear translation models. Each translation model is trained using the same data on which different segmenters are applied on the Arabic side (the *moses* decoder can handle multiple translation tables). So, the choice of the best segmentation is done during decoding, simultaneously to the construction of the best translation.

A very preliminary experiment conducted on 100 sentences from IWSLT07 and IWSLT08 data sets, for which we clearly observed differences of performance between both segmentations, lead to 35.15% BLEU for the multiple segmentation approach compared to 29.75% for the Buckwalter-based system and 25.36 for the ASVM-based system. These first results were obtained on translation inputs particularly well suited for our approach. More reliable (and unfortunately disappointing!) experiments are provided in section 5 of this paper.

However, *Figure 2* below illustrates an interesting aspect of our technique: it shows how an "hybrid" segmentation path can be chosen during decoding. On this figure, the top path corresponds to Buckwalter segmentation (MT output of such a chain gives "*I'd like this car for about a week*") ; the chain below corresponds to ASVM segmentation (MT output gives "*to rent this car for long old almost*"). The plain line path corresponds to the segmentation chosen during the CN decoding process (MT output is "*I'd like to rent this car for about a week*"). The translation obtained in this latter case is the best and would have never been obtained without using a multiple segmentation approach.

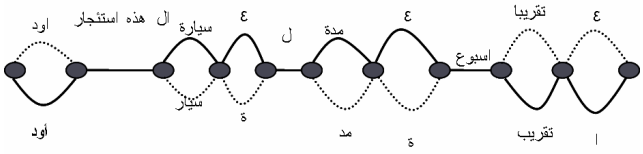

Fig. 2. Translation of an Arabic sentence using two different segmentations (top: Buckwalter - below: ASVM – CN decoding path in plain line)

4. Multiple segmentation for spoken language translation[5]

[4] We are aware that the last version of *moses* now allows to decode lattices but this feature was not available yet when this work was done.

[5] Unfortunately, this year at IWLST, the AE task is no longer a spoken language translation task since no ASR output was provided. However, we conducted spoken language translation experiments in this work using 2006 to 2008 data sets.

In this section, we describe how the multiple segmentation approach presented before can be applied to the spoken language translation case where ASR output lattices are provided to the MT system input.

4.1. Lattice decomposition for spoken language translation

In spoken language translation, one problem we sometimes face is that the word graphs provided by the ASR system do not have necessarily word decomposition compatible with the word decomposition used to train our MT models. It is actually the case in the framework of IWSLT evaluations where the Arabic ASR system used to generate the lattices is unknown to the participants. In order to handle this problem, we have already proposed a word lattice decomposition process to make the lattices (and then the word CN) compatible with our own level of decomposition. This process is described more precisely in [4] and [6]. In a few words, the decomposition algorithm implements the following steps:

1. Based on a word/sub-word dictionary or a morphological segmenter, all decompoundable words in the word lattice are identified.
2. Each of these words is decomposed into a sequence of sub-words that depends on the number of sub-words in the word. Some new nodes and links are then inserted in the word lattice.
3. For each new decomposed sub-word in the current word lattice, the new acoustic score and the duration are modified: the duration and the acoustic scores of the initial word are proportionally divided into sub-words duration and scores as a function of the number of graphemes in the sub-words.
4. An approximation is made for the LM score: the LM score corresponding to the first sub-word of the decomposed word is equal to the LM score of the initial word, while we assume that after the first sub-word, there is only one path to the last sub-word of the word (so the following LM scores are made equal to 0).
5. Finally, the new subword lattice is converted into a CN using an algorithm similar to [7].

4.2. Multiple segmentation for spoken language translation

The multiple segmentation process for spoken language translation is presented in *Figure 3*. The ASR lattice (marked as a "word lattice" in the figure, but it is more accurate to say that it is a lattice made up of "unknown" units) is decomposed according to the different sub-word sets (corresponding to different morphological segmentations). Then we create a new starting node S and a new ending node E for the common lattice. We link the node S with starting nodes of all subword lattices (n°1 and n°2) and link ending nodes of all lattices with E. After this step, all lattices are merged into a common lattice. This operation can also be seen as a "union" of lattices [8]. Finally, the obtained lattice is converted into a CN which will keep both ASR ambiguity and Arabic segmentation ambiguity. This latter CN is the input of the translation system which uses, as in section 3, multiple phrase tables corresponding to multiple Arabic segmenters.

5. Multiple Segmentation Experiments

5.1. Tools and data used

Since 2007, the LIG laboratory participates yearly to the IWSLT evaluation campaign (Arabic – English speech translation task). In the experiments reported here, we have used the data provided by the IWSLT09 organizers and a few publicly available additional data.

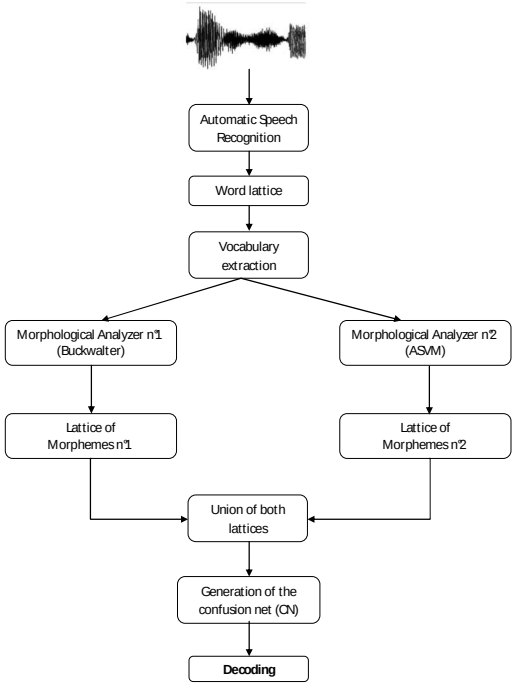

Fig. 3. Multiple segmentation process for spoken language translation

To train the translation models, the *train* part of the IWSLT09 data was used (a training corpus of 19972 sentence pairs). As development data, we used several subsets provided: the *dev4* subset, made up of 489 sentences, which corresponds to the IWSLT06 development data (we will refer, in the rest of the paper, to *dev06* for this data set); the *dev5* subset, made up of 500 sentences, which corresponds to the IWSLT06 evaluation data (we will refer, in the rest of the paper, to *tst06* for this data set); and the *dev6* subset, made up of 500 sentences, which corresponds to the IWSLT07 evaluation data (we will refer, in the rest of the paper, to *tst07* for this data set). The tuning of the MT model parameters (minimum error rate training) was systematically done on the *dev06* subset.

As additional data, we first used an Arabic / English bilingual dictionary of around 84k entries. This dictionary can be found online[6]. For English LM training, we also used out-of-domain corpora taken from the LDC's Gigaword corpus[7].

Our baseline speech translation system was built using tools available in the MT community:

- GIZA++ [9] was used for the alignments,
- The *moses* [8] decoder (and the training / testing scripts associated) was used (2008-07-11 release),
- SRILM [10] was used to train the LMs and to deal with ASR word graphs,
- The Buckwalter morphological segmenter[9] and ASVM (a free

[6] http://freedict.cvs.sourceforge.net/freedict/eng-ara/
[7] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05
[8] *Moses* open source project: http://www.statmt.org/moses
[9] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L49

Arabic segmenter developed at Columbia University[10]) were used for Arabic word segmentation,

- All the performances reported in this paper are BLEU[11] [11], NIST [12] and METEOR [13].

5.2 Baseline systems

Our systems were trained on the 20k *train* bitext provided concatenated to the bilingual dictionary of 84k entries, described in the previous section. The *moses* training script (default options) was used to build a phrase translation table from the bitext. The Arabic part of the bitext was systematically segmented using both Buckwalter and ASVM morphological segmenters to train two different phrase tables. On the English side, we removed punctuation and case (both pieces of information are further restored after translation using *hidden-ngram* and *disambig* from the SRILM toolkit [10]). The weights (5 for translation model, 1 for the language model, 1 for distorsion and 1 for word penalty) are optimized by means of a minimum error training (MERT) procedure. The default distortion limit is set to 6.

For English language modeling, we used both in-domain (English part of the *train* bitext) and out-of-domain (LDC's Gigaword corpus) to train the English LM. The interpolation weights (0.7/0.3) optimize the perplexity on the *dev06* corpus. The default options of the moses decoder are used and unknown words are dropped from the translation output. When an ASR output is provided for translation, CN decoding is performed as explained in section 4. Note that all the parameters of the log-linear model used for the CN decoder are systematically retuned on dev06 set (since an additional parameter, corresponding to the CN posterior probability is added in that case, as described in [5]). More details on these baseline systems can be found in [4] and [6].

5.3 Experiments and results

The text and speech translation performance are reported in *table 2* and *table 3* respectively. In these tables, we show results for different Arabic segmenters as well as for the multiple segmentation approach described in sections 3 and 4.

Table 2
Text (verbatim) translation results for different Arabic segmenters as well as for the multiple segmentation approach

	Dev06			Tst06			Tst07		
Score	BLEU	NIST	METEOR	BLEU	NIST	METEOR	BLEU	NIST	METEOR
ASVM	32.34	7.07	67.46	25.47	6.37	64.87	50.52	8.18	**77.93**
Buckwalter	**35.11**	**7.47**	**68.72**	28.17	**6.84**	65.61	**50.79**	**8.27**	77.56
Multiple	34.39	7.28	64.24	**29.19**	6.72	**67.31**	48.47	7.99	77.32

Table 3
Speech (ASR lattices) translation results for different Arabic segmenters as well as for the multiple segmentation approach

	Dev06			Tst06			Tst07		
Score	BLEU	NIST	METEOR	BLEU	NIST	METEOR	BLEU	NIST	METEOR
ASVM	28.27	**6.47**	62.92	22.78	5.91	60.99	41.04	7.13	70.92
Buckwalter	27.69	6.40	63.04	**24.19**	**6.09**	62.14	42.54	7.29	71.44
Multiple	**29.27**	5.57	**66.73**	24.01	5.21	**66.09**	**42.72**	**7.41**	**72.60**

5.4 Discussion

Concerning the text translation results, we observe, on *tst06*, an improvement of BLEU and METEOR scores using the multiple segmentation approach compared to the best Arabic segmentation used alone (Buckwalter). However, this improvement is not significant and is not observed on *tst07* corpus. Actually, after analyzing more deeply the translation outputs, we noticed that the multiple segmentation approach can introduce errors. For instance, *figure 4* shows a wrong segmentation path chosen during decoding. While both ASVM and Buckwalter segmentation lead to the same correct translation (*"Can I use your phone?"*), the multiple segmentation approach translates the CN as *"I use your phone"*. It seems that one drawback of the approach is that it tends to favor short paths in the CN (and consequently short translations). This problem is more important on *tst07* which contains utterances significantly shorter than in *tst06* and *dev06*. We did not deal with this problem yet but one possibility would be to add probabilities on the CN transitions in order to penalize short Arabic segments.

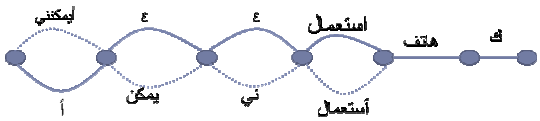

Fig. 4. Example of error: translation of an Arabic sentence using two different segmentations (top: Buckwalter - below: ASVM - CN decoding path in plain line).

Concerning the speech translation results, BLEU and METEOR are better using our multiple segmentation technique. If we analyze the results globally on all test sets (*dev06*, *tst06*, *tst07*) where ASVM is sometimes better than Buckwalter and sometime worse, the multiple segmentation technique seems to be efficient to improve the general system performance. Finally, *Table 4* shows a few translations obtained with Buckwalter, ASVM and multiple segmentation approaches.

Table 4
Examples of English translations obtained with Buckwalter, ASVM and multiple segmentations

Buckwalter	How will you pay in cash or card
ASVM	How do you pay in cash credit card
Multiple	**How will you pay in cash or credit card**
Buckwalter	I'd like this car for about a week
ASVM	To rent this car for long old almost
Multiple	**I'd like to rent this car for about a week**
Buckwalter	I'm sorry sir non-smoking a on the train
ASVM	Sorry sir non-smoking a on the train
Multiple	**I'm sorry sir non-smoking seat on the train**

6. LIG submission for IWSLT09

The 2009 LIG submission is based on the work presented in the previous sections. However, since a release of the moses decoder was provided in April 2009 (2009-04-13), we performed a comparison of both versions of moses before taking a decision on which system should be submitted as the primary system. *Table* 5 presents this comparison between moses 2008 and 2009 versions[12].

[10] http://www1.cs.columbia.edu/~mdiab/
[11] Mt-eval v12 is used
[12] It is important to note that in these experiments, the moses decoder and the training scripts (used to train the phrase table) are different.

Table 5
Comparison of text and speech translation results (BLEU) for moses 2008 and moses 2009 decoder and tools.

	dev06		tst06		tst07	
	verbatim	ASR	verbatim	ASR	verbatim	ASR
ASVM (moses 2008)	32.34	28.27	25.47	22.78	50.52	41.04
ASVM (moses 2009)	**35.96**	28.60	**29.7**	23.60	**51.14**	40.61
Buckwalter (moses 2008)	35.11	27.69	28.17	**24.19**	50.79	42.54
Buckwalter (moses 2009)	34.02	27.43	27.79	23.89	48.89	42.03
Multiple (moses 2008)	34.39	**29.27**	29.19	24.01	48.47	42.72
Multiple (moses 2009)	32.11	28.22	26.93	24.05	45.51	**42.80**

Based on these results, and since IWSLT09 AE task was focused on verbatim transcription only, we decided to submit our ASVM-based MT system, using moses 2009, as *primary* system. The "Multiple segmentation" and "Buckwalter" systems were submitted as systems *contrastive1* and *contrastive2* respectively. The preliminary automatic evaluation results obtained by the LIG at IWSLT09 are presented in *table 6*.

Table 6
Preliminary automatic evaluation results obtained by the LIG at IWSLT09

	BLEU	NIST	METEOR
LIGprimary	**46.62**	**7.86**	**73.69**
LIGcontrastive1	43.13	7.40	70.04
LIGcontrastive2	44.35	7.72	72.68

Finally, *figure 5* presents the evaluation of the LIG AE system measured by running and evaluating our 2007, 2008 and 2009 systems on the same data sets (dev06, tst06 and tst07). The results show a yearly improvement of our AE MT system.

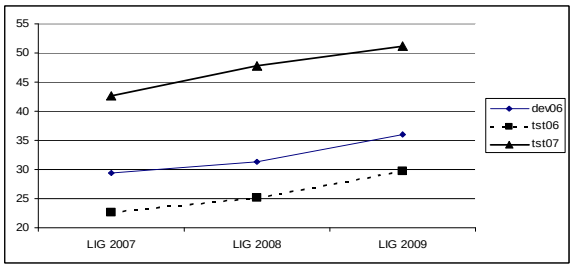

Figure 5 Evaluation of the verbatim text translation performance (BLEU) for LIG systems from 2007.

7. Conclusion

In this work, we were interested in Arabic to English Statistical Machine Translation (SMT). Arabic is a morphologically rich language, and recent experimentations in our laboratory have shown that the performance of Arabic to English SMT systems varies greatly according to the Arabic morphological segmenters applied. Based on this observation, we proposed to use simultaneously multiple segmentations for machine translation of Arabic. The core idea is to keep the ambiguity of the Arabic segmentation at the system input (using confusion networks). Then, we hope that the best segmentation will be chosen during MT decoding. The mathematics of this multiple segmentation approach was given. Practical implementations in the case of verbatim text translation as well as speech translation were proposed. Experiments conducted in the framework of IWSLT evaluation campaign have shown the potential of the multiple segmentation approach for spoken language translation.

The 3 systems (ASVM, Buckwalter and Multiple) described in this paper were presented at IWSLT09 evaluation and details on the LIG submission were also given in this article (section 6). The problem of short sentences mentioned in *section 5* will be addressed in future works. One other problem is the fact that CN representation introduces new segmentation paths that are incorrect. A true lattice might be better to represent the segmentation ambiguity instead of a CN and it could be experimented in the future since *moses* decoder was recently released to decode such lattice structures.

8. References

[1] M. Maxim Khalilov & al. The TALP&I2R SMT Systems for IWSLT 2008. IWSLT08. Hawaï, USA. 2008.

[2] W. Shen & al. The MIT-LL/AFRL IWSLT-2008 MT System. IWSLT08. Hawaï, USA. 2008.

[3] M. Afify, R. Sarikaya, H-K. J. Kuo, L. Besacier, and Y. Gao "On the use of morphological analysis for dialectal Arabic speech recognition", in Proc. ICSLP'06, Pittsburgh, USA, 2006.

[4] L. Besacier, A. Ben-Youcef, H. Blanchon « The LIG Arabic / English Speech Translation System à IWSLT08 » IWSLT08. Hawai. USA. October 2008.

[5] Bertoldi, N., Zens, R. and Federico, M., "Speech Translation by Confusion Network Decoding", ICASSP'07, vol. 4, pp. 1297-1300, Honolulu, Hawaii, April 2007.

[6] L. Besacier, A. Mahdhaoui, V-B Le, « The LIG Arabic / English Speech Translation System at IWSLT07 » IWSLT07. Trento. Italy. October 2007.

[7] Mangu, L., Brill, E., and Stolcke, A., "Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks", Computer Speech and Language, vol. 14, no. 4, pp. 373-400, 2000.

[8] M. Mohri, "Finite-State Transducers in Language and Speech Processing", Computational Linguistics, vol. 23, no. 2, pp. 269-311, 1997.

[9] Och, F. J. and Ney, H., "A Systematic Comparison of Various Statistical Alignment Models ", Computational Linguistics, vol. 29, no. 1, pp. 19-51, March 2003.

[10] Stolcke, A., "SRILM - An Extensible Language Modeling Toolkit", ICSLP'02, vol. 2, pp. 901-904, Denver, Colorado, September 2002.

[11] Papineni, K., Roukos, S., Ward, T., and Zhu, W., "BLEU: A method for automatic evaluation of machine translation", ACL'02, pp. 311-318, Philadelphia, USA, July 2002.

[12] Doddington, G. (2002) Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Proc. HLT 2002. San Diego, California. March 24-27, 2002. vol. 1/1: pp. 128-132

[13] Lavie, A., A. Agarwal. "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments", Proceedings of Workshop on Statistical Machine Translation at the 45th Annual Meeting of the ACL (ACL-2007), Prague, June 2007.