

Mult-Engine System for the 2008 NIST MT Evaluation

1.0 SITE AFFILIATION

CASIA - Institute of Automation, Chinese Academy of Sciences

2.0 CONTACT INFORMATION

Jinhua Du - jhdu@hitic.ia.ac.cn
Wei Wei - weiwei@hitic.ia.ac.cn
Zhendong Yang - zdyang@hitic.ia.ac.cn
Wei Pang - wpang@hitic.ia.ac.cn
Weihua Wang - whwang@hitic.ia.ac.cn
Bo Xu - xubo@hitic.ia.ac.cn

3.0 SUBMISSIONS:

CASIA_chinese_constrained_primary
CASIA_chinese_constrained_contrast1
CASIA_chinese_constrained_contrast4
CASIA_chinese_constrained_contrast19

4.0 PRIMARY SYSTEM SPECS:

Our primary system is a multiple combination system, including the follows:

- 1) a phrase-based system with zero fertility expansion
- 2) a hierarchical phrase-based system

A cascaded framework for statistical machine translation system combination is proposed. The framework integrates the MBR and word alignments techniques to process the outputs of multiple machine translation systems. It is easy for the proposed framework to combine different systems like phrase-based system, syntax-based systems and so on. Our framework includes the following three steps:

- 1) Use MBR decoder to select a best hypothesis as the alignment reference from the outputs of multiple translation systems.
- 2) Use the alignment reference to perform the word alignment with the other hypotheses by GIZA++ toolkit[1], and then build the word transition network based on our modified TER scheme.
- 3) Use beam search decoding to search the best translation from the constructed WTN.

4.1 CORE MT ENGINE ALGORITHMIC APPROACH

- 1) a system combination framework of MBR decoding and confusion network decoding [2]

- 2) a hierarchical phrase-based engine [3]
- 3) a phrase-based engine
- 4) a syntax augmented machine translation engine[4]
- 4) a string-to-tree engine

4.2 CRITICAL ADDITIONAL FEATURES AND TOOLS USED

- 1) ICT-CAS Chinese word segment system [5]
- 2) Adwait Ratnaparkhi's Part-Of-Speech Tagging tool [6]
- 3) GIZA++

4.3 SIGNIFICANT DATA PRE/POST-PROCESSING

1)Parallel data

LDC2000T46

LDC2000T50

LDC2002E18

LDC2002E27

LDC2002L27

LDC2002T01

LDC2003E07

LDC2003E14

LDC2003T17

LDC2004E12

LDC2004T07

LDC2004T08

LDC2005T01

LDC2005T06

LDC2005T10

LDC2005T34

LDC2006T04

LDC2007T09

2)Monolingual data

LDC2007T07

5.0 KEY DIFFERENCE IN CONTRASTIVE SYSTEMS

5.1 Hierarchical phrase-based system(Contrast1)

The CASIA_chinese_constrained_contrast1-3 are CKY-style hierarchical phrase-based systems built on a synchronous context-free grammar rules. The core algorithm of the decoder is borrowed from the CKY chart based parsing algorithm. During decoding, the source sentence is annotated with word alignments using the method of Koehn et al [7], which combines the GIZA++ results of both directions based on heuristics. Then the SCFG rules $X \rightarrow (r, a, \sim)$ is used to

translate source phrase r into target phrase a . These rules will be used continuously until the whole source sentence is covered.

The key differences among 3 submissions are parameters trained on different development sets.

5.2 Phrase-based system with zero fertility expansion(Contrast4)

Our system applies a phrase -based translation model to capture the corresponding relationship between two languages. We learn the phrase alignments from a corpus that the words are aligned by a training toolkit for a word-based translation model: the Giza++ toolkit[8] for the IBM models[9]. The extraction heuristic is similar to the one used in the alignment template work by Och et al[10]. The phrase-based decoder we developed employs a beam search algorithm, similar to the one in [7], but it applies the words with fertility probability of zero in the target language [11].

5.3 String-to-tree system(Contrast19)

we use chinese string and english phrasing information in this system,which uses string to tree template to translate, we combine some hierarchical phrases in translation.

6.0 REFERENCES:

- [1] F. J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51
- [2] Jinhua Du, Wei Wei, Zhendong Yang and Bo Xu: A Cascaded Framework for Statistical Machine Translation System Combination. In proceedings of Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering, 2007
- [3] Wei Wei ,Bo Xu: Hierarchical chunking-phrase based translation. In proceedings of Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering, 2007
- [4] Andreas Zollmann and Ashish Venugopal. 2006. Syntax Augmented Machine Translation via Chart Parsing. In NAACL 2006 - Workshop on statistical machine translation, New York. June 4-9.
- [5] ZHANG Hua-Ping et al, Chinese Lexical Analysis Using Hierarchical Hidden Markov Model , Second SIGHAN workshop affiliated with 41th ACL; Sapporo Japan, July, 2003, pp. 63-70
- [6] Ratnaparkhi A. A maximum entropy model for part-of-speech tagging. In: Brill E, Church K, eds. *Proc. of the Conf. on Empirical Methods in Natural Language Processing*. Somerset: Association for Computational Linguistics, 1996. 133-142.
- [7] Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pp. 127-133
- [8] Franz Josef Och and Hermann Ney (2000). Improved Statistical Alignment Model. In *Proceeding of the 38th Annual Meeting of the ACL*, pages 440-447.
- [9] Peter. F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, *The Mathematics of Statistical Machine Translation: Parameter Estimation*, Computational Linguistics, Vol 19, No.2 ,1993
- [10] Franz Josef Och, Tillmann, C., and Hermann Ney.(1999). improved alignment models for statistical machine translation. In proceedings of the Joint Conference of Empirical Methods in

Natural Language Processing and Very Large Corpora, pp.20-28.

[11] Wei Wei, Wei Pang, Zhendong Yang, Zhenbiao Chen, Chengqing Zong, Bo Xu. 2006. CASIA SMT System For TC-STAR Evaluation Campaign 2006. In Proceedings of TC-STAR Workshop on Speech-to-Speech Translation, pages 69-74, Barcelona, Spain, June.