# System Combination with Extra Alignment Information

*Xiaofeng Wu Tsyoshi Okita Josef van Genabith Qun Liu*
Centre of Next Generation Localisation(CNGL), School of Computing, Dublin City University
{xiaofengwu,tokita,josef,qliu}@computing.dcu.ie

ABSTRACT
This paper provides the system description of the IHMM team of Dublin City University for our participation in the system combination task in the Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT (ML4HMT-12). Our work is based on a confusion network-based approach to system combination. We propose a new method to build a confusion network for this: (1) incorporate extra alignment information extracted from given meta data, treating them as sure alignments, into the results from IHMM, and (2) decode together with this information. We also heuristically set one of the system outputs as the default backbone. Our results show that this backbone, which is the RBMT system output, achieves an 0.11% improvement in BLEU over the backbone chosen by TER, while the extra information we added in the decoding part does not improve the results.

KEYWORDS: system combination, confusion network, indirect HMM alignment, backbone chosen.

*Second ML4HMT Workshop*, pages 37–44,
COLING 2012, Mumbai, December 2012.

## 1 Introduction

This paper describes a new extension to our system combination module in Dublin City University for the participation in the system combination task in the ML4HMT-2012 workshop. We incorporate alignment meta information to the alignment module when building a confusion network.

Given multiple translation outputs, a system combination strategy aims at finding the best translations, either by choosing one sentence or generating a new translation from fragments originated from individual systems(Banerjee et al., 2010). Combination methods have been widely used in fields such as parsing (Henderson and Brill, 1999) and speech recognition (Fiscus, 1997). In late the 90s, the speech recognition community produced a confusion network-based system combination approach, spreading instantly to SMT community as well.

The traditional system combination approach employs confusion networks which are built by the monolingual alignment which induces sentence similarity. Confusion networks are compact graph-based structures representing multiple hypothesises (Bangalore et al., 2001). It is noted that there are several generalized forms of confusion networks as well. One is a lattice (Feng et al., 2009) and the other is a translation forest (Watanabe and Sumita, 2011). The former employs lattices that can describe arbitrary mappings in hypothesis alignments. A lattice is more general than a confusion network. By contrast, a confusion forest exploits syntactic similarity between individual outputs.

Up to now, various state-of-the-art alignment methods have been developed including Indirect-HMM (He et al., 2008; Du and Way, 2010) which is a statistical-model-based method, and TER which is a metric-based method which uses an edit distance. In this work we focus on the IHMM method.

The main problem of IHMM is that there are numerous one-to-many and one-to-null cases in the alignment results. This alignment noise significantly affects the confusion network construction and the decoding process. In this work, in addition to the IHMM alignment, we also incorporate alignment meta information extracted from an RBMT system to help the decoding process.

The other crucial factor is the backbone selection which also affects the combination results. The backbone determines the word order in the final output. Backbone selection is often done by Minimum Bayes Risk (MBR) decoding which selects a hypothesis with minimum average distance among all hypotheses (Rosti et al., 2007a,b). In this work we heuristically choose an RBMT output as the backbone due to its (expected) overall grammatically well-formed output and better human evaluation results.

We report our results and provide a comparison with traditional confusion-network-based network approach.

The remainder of the paper is organized as follows: We will review the state-of-the-art system combination framework based on confusion networks in Section 2. We describe our experimental setup, how we extract the alignment information from meta-data and how we use it in Section 3. The results and analysis are also given in this section. We draw conclusions in Section 4.

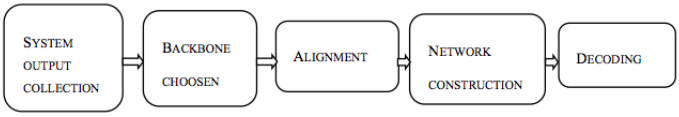


Figure 1: Word level confusion network based system combination architecture

$H_1$: He jogs in the park
$H_2$: He runs the park
$H_3$: Garden he runs
$H_4$: It in garden

(a)

| | | | | | |
|---|---|---|---|---|---|
| He | jogs | in | ε | the | park |
| He | runs | ε | ε | the | park |
| he | runs | ε | ε | ε | Garden |
| it | ε | in | ε | ε | garden |

(b)

Figure 2: a) The hypothesises collected; (b) the final confusion net work constructed.

## 2 Background on Confusion Networks

### 2.1 Confusion Network Architecture

The current state-of-art approach to word level system combination is described in (Rosti et al., 2007b). The system architecture is illustrated in Figure 1.

Suppose we have collected four system outputs $H_1$-$H_4$ which are shown in Figure 2a. Then the traditional way of choosing a backbone is to use minimum average edit distance (or other measurements) as shown in Equation 1.

$$B = H^* = argmin_{H \in \nabla} \sum_{H \in \nabla} (H_i, H) \tag{1}$$

The backbone is used to decide the word order of the final output. After obtaining the backbone, all other hypotheses are aligned to it. The alignment strategies include IHMM, TER, etc. Note that during the word alignment word reordering and 'null' insertion are performed, which is usually called normalization. The confusion network, which can be constructed directly from the normalized alignment is given in Figure 2b, in which case $H_1$ is chosen as the backbone.

### 2.2 Indirect HMM Alignment

In this work we implement the IHMM (He et al., 2008). IHMM is a refined version of HMM alignment (Vogel et al., 1996) which is widely used in bilingual word alignment (Och and Ney, 2003).

Let $B = (b_1, ..., b_J)$ denote the $J$ words in the backbone sentence, $H = (h_1, ..., h_I)$ denote one of the hypothesis, and $A = (a_1, ..., a_J)$ denote the alignment of each backbone word to the hypothesis word. We use Equation 2 to compute the alignment probability of each word pair. In Equation 2, $d$ represents the distortion model and $p$ denotes the word similarity model.

$$P(H|B) = \sum_{A} \prod_{j=1...J} [d(a_j|a_{j-1}, I) p(h_j|b_{a_j})] \tag{2}$$

In order to handle the words which are aligned to an empty word, we also insert a *null* associated with each backbone word.

We follow (Vogel et al., 1996) and use Equation 3 to compute the distortion model.

$$d(i|i',I) = \frac{c(i-i')}{\sum_{l=1}^{I} c(l-i)} \tag{3}$$

Where $c(\Delta)$ represents the word distance grouped into $c(\leq -4), c(-3), ..., c(0), ..., c(5), c(\geq 6)$ 13 buckets, and computed with Equation 4 which peak at $\Delta = 1$.

$$c(\Delta) = (1 + |\Delta - 1|)^{-2} \tag{4}$$

The word similarity probability in Equation 2 is computed by Equation 5. We use two small Chinese-English & English-Chinese dictionaries (10k entries each) to compute $p_{semantic}$, and the longest common subsequence matching score to obtain $p_{surface}$.

$$p(h_j|b_i) = \alpha p_{semantic}(h_j|b_i) + (1-\alpha) p_{surface}(h_j|b_i) \tag{5}$$

Given an HMM model, we use the Viterbi algorithm to obtain the one-to-many alignment, and by reordering and inserting *null* to their proper position both in the backbone and hypothesis, the final normalized alignment are produced, as shown in Figure 2b.

## 2.3 Decoding & Parameter tuning

We use a log linear combination strategy shown in Equation 6, which is described in (Rosti et al., 2007a), to compute the hypothesis confidence.

$$logp(E|F) = \sum_{i=1}^{N_{nodes}-1} log(\sum_{i=1}^{N_{system}} w_l p(word|l,i)) + \nu Lm(E) + \mu N_{nulls}(E) + \epsilon Len(E) \tag{6}$$

where $N_{nodes}$ is the number of nodes the current confusion network has, $N_{system}$ is the number of sytems, $w$ denotes the system weight, $Lm$ represents the language model score of the current path, $N_{nulls}$ stands for the number of nulls inserted, and Len is the length of the current path. $\nu$ ,$\mu$ and $\epsilon$ are the corresponding weights of each feature.

A beam search algorithm is employed to find the best path.

# 3 Experimental Setup

## 3.1 Data

We participate in the ML4HMT-12 shared task ES-EN. Participants are given a development bilingual data set aligned at the sentence level. Each "bilingual sentence" contains: 1) the source sentence, 2) the target (reference) sentence and 3) the corresponding multiple output translations from four systems, based on different MT approaches (Ramırez-Sánchez et al., 2006; Alonso and Thurmair, 2003; Koehn et al., 2007). The output has been annotated with

system-internal meta-data information derived from the translation process of each of the systems.

In this work we use 1000 sentence pair from the 10K development set to tune the system parameters and all the 3003 sentence pairs in the test set to run the test.

## 3.2 Backbone Selection

Equation 1 describes the traditional backbone selection. However in this work we heuristically set Lucy RBMT (Alonso and Thurmair, 2003) output as the backbone. Our motivations are that: 1) Lucy's output tends to be more grammatical than Moses or other MT systems; 2) according to the previous ML4HMT-2011, Lucy has better human evaluation scores than other statistical machine translation systems.

## 3.3 Alignment Extraction of Lucy

The Lucy LT RBMT (Alonso and Thurmair, 2003) takes three steps to translate a source language string into a target language string: an analysis step, a transfer step, and a generation step. The meta data annotations provided in ML4HMT development set follows these three steps describing the parse tree for the respective translation steps.

We extract the alignment from this annotation in the following manner. First, we extract tuples where each connects a source word, intermediate words, and a target word by looking at the annotation file. There are some alignments dropped in this process. Such dropped alignments include the alignment of UNK (unknown)words marked by the Lucy LT RBMT system, words such as "do" which will not appear in the transfer step, and so forth. One remark is that since we trace these annotations based on the parse tree structure provided by the Lucy LT RBMT system, the exact order in the sentence is sometimes lost. This caused a problem when there are multiple "the", "of", and so forth, tokens in the string, so that for a given "the" in sources there are multiple target position. Second, we delete the intermediate representations, and obtain the source and the target words/phrases pairs.

Examples of the extracted alignments (the second sentence in test set) are shown in Figure 3.

From Figure 3 we can see that words like 'últimos' which has a one-to-one alignment are all correct alignments, while words like 'de' or 'el' which are involved in many-to-many alignment, carry much less confidence for the alignment. Given this observation, one idea would be using these extracted sure alignments, which are one-to-one, to guide decoding.

## 3.4 Decoding with Alignment Bias

In the decoding part, we change the Equation 6 into Equation 7 as follows

$$p(E_\psi) = \theta_\psi log p(E_\psi|F) \tag{7}$$

where $\psi = 1...N_{nodes}$ denotes the current node at which the beam search arrived, and $\theta_\psi = 1$ if a current node is not a sure alignment extracted from Lucy's meta-data and $\theta_\psi > 1$, otherwise.

```
TGT:the(0) period(1) aktuálně.cz(2) "(3) examined(4) "(5) the(6) members(7) of(8) the(9)
new(10) board(11) of(12) the(13) čssd(14) to(15) check(16) its(17) knowledge(18) of(19) the(20)
language(21) marked(22) slang(23) that(24) has(25) risen(26) up(27) in(28) the(29) last(30)
years(31) in(32) the(33) board(34) ,(35) when(36) the(37) current(38) members(39) of(40) the(41)
coalition(42)  governed(43) prague(44) .(45)
SRC:El(0) período(1) Aktuálně.cz(2) "(3) examinó(4) "(5) a(6) los(7) miembros(8) del(9) nuevo(10)
Consejo(11) del(12) ČSSD(13) para(14) comprobar(15) sus(16) conocimientos(17) del(18) lengua(19)
marcado(20) slang(21) que(22) ha(23) surgido(24) en(25) los(26) últimos(27) años(28) en(29) el(30)
Consejo(31) ,(32) cuando(33) gobernaban(34) Praga(35) los(36) actuales(37) miembros(38) de(39)
la(40) coalición(41) .(42)
2 ||| [[u'el'], [[30]], [u'the'], [[0, 6, 9, 13, 20, 29, 33, 37, 41]]]
2 ||| [[u'per\xedodo'], [[1]], [u'period'], [[1]]]
2 ||| [[u'examin\xf3'], [[4]], [u'examined'], [[4]]]
2 ||| [[u'a'], [[6]], [u'to'], [[15]]]
2 ||| [[u'los'], [[7, 26, 36]], [u'the'], [[0, 6, 9, 13, 20, 29, 33, 37, 41]]]
2 ||| [[u'miembros'], [[8, 38]], [u'members'], [[7, 39]]]
2 ||| [[u'de'], [[39]], [u'of'], [[8, 12, 19, 40]]]
2 ||| [[u'l'], ['-1'], [u'the'], [[0, 6, 9, 13, 20, 29, 33, 37, 41]]]
2 ||| [[u'nuevo'], [[10]], [u'new'], [[10]]]
2 ||| [[u'consejo'], [[11, 31]], [u'Board'], [[11, 34]]]
2 ||| [[u'comprobar'], [[15]], [u'check'], [[16]]]
2 ||| [[u'sus'], [[16]], [u'its'], [[17]]]
2 ||| [[u'conocimientos'], [[17]], [u'knowledge'], [[18]]]
2 ||| [[u'lengua'], [[19]], [u'language'], [[21]]]
2 ||| [[u'surgido'], [[24]], [u'risen'], [[26]]]
2 ||| [[u'en'], [[25, 29]], [u'in'], [[28, 32]]]
2 ||| [[u'\xfaltimos'], [[27]], [u'last'], [[30]]]
2 ||| [[u'a\xf1os'], [[28]], [u'years'], [[31]]]
2 ||| [[u'cuando'], [[33]], [u'when'], [[36]]]
2 ||| [[u'gobernaban'], [[34]], [u'governed'], [[43]]]
2 ||| [[u'Praga'], [[35]], [u'Prague'], [[44]]]
2 ||| [[u'actuales'], [[37]], [u'current'], [[38]]]
2 ||| [[u'la'], [[40]], [u'the'], [[0, 6, 9, 13, 20, 29, 33, 37, 41]]]
2 ||| [[u'coalici\xf3n'], [[41]], [u'coalition'], [[42]]]
```

Figure 3: extracted alignment from Lucy LT RBMT meta data.

### 3.5 Experimental Results and Analysis

In our experiments, we set $\alpha$ 0.1 in Equation 5, according to (Feng et al., 2009). All development set and test set data are tokenized and lower cased. We use mteval-v13.pl[1], no-smoothing and case sensitive for the evaluation.

Table 1 shows the result of using Lucy as a backbone and the result of changing $\theta_\psi$ on the development and test sets. Note that $\theta_\psi = 1$ stands for the case when there is no effect of this factor on the current path.

| $\theta_\psi$ | Devset(1000) | | Testset(3003) | |
|---|---|---|---|---|
| | NIST | BLEU | NIST | BLEU |
| 1 | 8.1328 | 0.3376 | 7.4546 | 0.2607 |
| 1.2 | 8.1179 | 0.3355 | 7.2109 | 0.2597 |
| 1.5 | 8.1171 | 0.3355 | 7.4512 | 0.2578 |
| 2 | 8.1252 | 0.3360 | 7.4532 | 0.2558 |
| 4 | 8.1180 | 0.3354 | 7.3540 | 0.2569 |
| 10 | 8.1190 | 0.3354 | 7.1026 | 0.2557 |

Table 1: The Lucy backbone with tuning of $\theta_\psi$.

From Table 1, we see a slight decrease of quanlity when we increased the factor. But an

[1]ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13.pl

interesting observation is that when we increased $\theta_\psi$ to 10 the result was not much affected. We believe this is since the sure alignments which we extracted from the Lucy alignments were almost perfectly consistent with the alignment resulting from IHMM. The best path derived by IHMM included most of the sure alignments extracted from Lucy.

| | Devset(1000) | | Testset(3003) | |
|---|---|---|---|---|
| TER Backbone | 8.1168 | 0.3351 | 7.1092 | 0.2596 |
| Lucy Backbone | 8.1328 | 0.3376 | 7.4546 | 0.2607 |

Table 2: TER Backbone selection results.

We compared results with thos obtained by Lucy backbone (which are in the Table 1 when $\theta_\psi = 1$) with that of the TER backbone in Table 2. We can see that the Lucy backbone result was 0.11% better than that of TER. This confirmed our assumption that Lucy would be a good backbone in system combination.

## 4 Conclusion

In this paper we describe new approaches we applied in building a confusion network. We focus on backbone selection and adding extra alignment information. Our results show that with choosing Lucy, which is an RBMT system, as a backbone the result is slightly better (0.11% improvment by BLEU) than the traditional TER backbone selection method. However the extra alignment information we added in the decoding part does not improve the performance. In our future work we will further analyse the reason for this.

## Acknowledgments

This work was supported by Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University.

## References

Alonso, J. and Thurmair, G. (2003). The comprendium translator system. In *Proceedings of the Ninth Machine Translation Summit*.

Banerjee, P., Du, J., Li, B., Kumar Naskar, S., Way, A., and Van Genabith, J. (2010). Combining multi-domain statistical machine translation models using automatic classifiers. Association for Machine Translation in the Americas.

Bangalore, B., Bordel, G., and Riccardi, G. (2001). Computing consensus translation from multiple machine translation systems. In *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*, pages 351–354. IEEE.

Du, J. and Way, A. (2010). Using terp to augment the system combination for smt. Association for Machine Translation in the Americas.

Feng, Y., Liu, Y., Mi, H., Liu, Q., and Lü, Y. (2009). Lattice-based system combination for statistical machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1105–1113. Association for Computational Linguistics.

Fiscus, J. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 347–354. IEEE.

He, X., Yang, M., Gao, J., Nguyen, P., and Moore, R. (2008). Indirect-hmm-based hypothesis alignment for combining outputs from machine translation systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 98–107. Association for Computational Linguistics.

Henderson, J. and Brill, E. (1999). Exploiting diversity in natural language processing: Combining parsers. In *Proceedings of the Fourth Conference on Empirical Methods in Natural Language Processing*, pages 187–194.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Annual meeting-association for computational linguistics*, volume 45, page 2.

Och, F. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Ramırez-Sánchez, G., Sánchez-Martınez, F., Ortiz-Rojas, S., Pérez-Ortiz, J., and Forcada, M. (2006). Opentrad apertium open-source machine translation system: an opportunity for business and research. In *Proceedings of the Twenty-Eighth International Conference on Translating and the Computer*. Citeseer.

Rosti, A., Ayan, N., Xiang, B., Matsoukas, S., Schwartz, R., and Dorr, B. (2007a). Combining outputs from multiple machine translation systems. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 228–235.

Rosti, A., Matsoukas, S., and Schwartz, R. (2007b). Improved word-level system combination for machine translation. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 312.

Vogel, S., Ney, H., and Tillmann, C. (1996). Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 836–841. Association for Computational Linguistics.

Watanabe, T. and Sumita, E. (2011). Machine translation system combination by confusion forest. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1249–1257. Association for Computational Linguistics.