CMU System Combination in WMT 2011

Kenneth Heafield and Alon Lavie
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA, USA
{heafield,alavie}@cs.cmu.edu

Abstract

This paper describes our submissions, `cmu-heafield-combo`, to the ten tracks of the 2011 Workshop on Machine Translation's system combination task. We show how the combination scheme operates by flexibly aligning system outputs then searching a space constructed from the alignments. Humans judged our combination the best on eight of ten tracks.

1 Introduction

We participated in all ten tracks of the 2011 Workshop on Machine Translation system combination task as `cmu-heafield-combo`. This uses a system combination scheme that builds on our prior work (Heafield and Lavie, 2010), especially with respect to language modeling and handling non-English languages. We present a summary of the system, describe improvements, list the data used (all of the constrained monolingual data), and present automatic results in anticipation of human evaluation by the workshop.

2 Our Combination Scheme

Given single-best outputs from each system, the scheme aligns system outputs then searches a space based on these alignments. The scheme is a continuation of our previous system (Heafield and Lavie, 2010) so we describe unchanged parts of the system in less detail, preferring instead to focus on new components.

2.1 Alignment

We run the METEOR matcher (Denkowski and Lavie, 2010) on every pair of system outputs for a given sentence. It identifies exact matches, identical stems (Porter, 2001) except for Czech, WordNet synonym matches for English (Fellbaum, 1998), and automatically extracted matches for all five target languages. The automatic matches come from pivoting (Bannard and Callison-Burch, 2005) on constrained data. An example METEOR alignment is shown in Figure 1, though it need not be monotone.

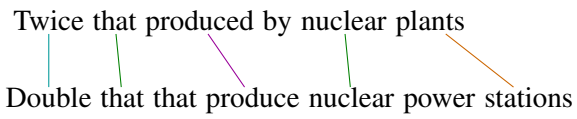


Figure 1: Alignment generated by METEOR showing exact (that–that and nuclear–nuclear), stem (produced–produce), synonym (twice–double), and unigram paraphrase (plants–stations) alignments.

2.2 Search

The search space is unchanged from Heafield and Lavie (2010), so we give a summary here. The general idea is to generate a combined sentence one word at a time, going from left to right. As the scheme creates an output, it also steps through the system outputs from left to right. Stepping through systems is synchronized with the partial output, so that words to the left are already captured in the hypothesis and the next word from any of the systems represents a meaningful extension of the partial output. All of these options are considered by hypothesis branching.

Thus far, we have assumed that system outputs are monotone: they agree on word order, so it is possible to step through all of them simultaneously. On the left are words captured in the partial output and on the right are the words whose meaning remains to be captured in the output. When systems disagree on word order, the partial output corresponds to disjoint pieces of a system's output. We still retain that notion that a word is either captured in the partial output or not captured, but do not have a single dividing line between them. In this case, we still proceed from left to right, considering the first uncaptured word for extension. Then, we skip over parts of a system's output that have already been captured.

Here, we have used the informal notion of words whose meaning is "captured" or "uncaptured" by the partial output. The system interprets words aligned to the partial output as captured while those not aligned to the hypothesis are considered uncaptured. A heuristic also cleans up excess words in order to keep the stepping process loosely synchronized across system outputs.

2.3 Features

We use three feature categories to guide search:

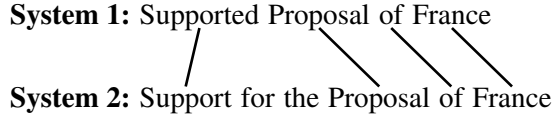
Length The length of the hypothesis in tokens.

Language Model Log probability and OOV count from an N-gram language model. Details are in Section 4.1.

Match Counts Counts of n-gram matches between systems outputs and the hypothesis.

The match count features report n-gram matches between each system and the hypothesis. Specifically, feature $m_{s,n}$ reports n-gram overlap between the hypothesis and system s. We track n-gram counts up to length N, typically 2 or 3, finding that tracking longer lengths adds little. An example is shown in Figure 2.

These match counts may be exact, in which case every word of the n-gram must be the same (up to case) or approximate, in which case any aligned word found by METEOR may be substituted. Because exact matches handle lexical choice and inexact matches collect more votes that better handle word order, we use both sets of features. However, the limit N may be different i.e. $N_e = 2$ counts exact matches up to length 2 and $N_a = 3$ counts inexact matches up to length 3.



Candidate: Support for Proposal of France

	Unigram	Bigram	Trigram
System 1	4	2	1
System 2	5	3	1

Figure 2: Example match feature values with two systems and matches up to length three. Here, "Supported" counts because it aligns with "Support".

3 Related Work

Hypothesis selection (Hildebrand and Vogel, 2009) selects an entire sentence at a time instead of picking and merging words. This makes the approach less flexible, in that it cannot synthesize new sentences, but also less risky by avoiding matching and related problems entirely.

While our alignment is based on METEOR, other techniques are based on TER (Snover et al., 2006), Inversion Transduction Grammars (Narsale, 2010), and other alignment methods. These use exact alignments and positional information to infer alignments, ignoring the content-based method used by METEOR. This means they might align content words to function words, while we never do. In practice, using both signals would likely work better.

Confusion networks (Rosti et al., 2010; Narsale, 2010) are the dominant method for system combination. These base their word order on one system, dubbed the backbone, and have all systems vote on editing the backbone. Word order is largely fixed to that of one system; by contrast, ours can piece together word orders taken from multiple systems. In a loose sense, our approach is a confusion network where the backbone is permitted to switch after each word.

Interestingly, BBN (Rosti et al., 2010) this year added a novel-bigram penalty that penalizes bigrams in the output if they do not appear in one of the sys-

tem outputs. This is the complement of our bigram match count features (and, since, we have a length feature, the same up to rearranging weights). However, they threshold it to indicate whether the bigram appears at all instead of how many systems support the bigram.

4 Resources

The resources we use are constrained to those provided for the shared task.

For the paraphrase matches described in Section 2.1, METEOR (Denkowski and Lavie, 2010) trains its paraphrase tables via pivoting (Bannard and Callison-Burch, 2005). The phrase tables are trained using parallel data from Europarl v6 (Koehn, 2005) (fr-en, es-en, de-en, and es-de), news commentary (fr-en, es-en, de-en, and cz-en), United Nations (fr-en and es-en), and CzEng (cz-en) (Bojar and Žabokrtský, 2009) sections 0–8.

4.1 Language Modeling

As with previous versions of the system, we use language model log probability as a feature to bias translations towards fluency. We add a second feature per language model that counts OOVs, allowing MERT to independently tune the OOV penalty. Language models often have poor OOV estimates for translation because they come not from new text in the same language but from new text in a different language. The distribution is even more biased in system combination, where most systems have already applied a language model. The new OOV feature replaces a previous feature that reported the average n-gram length matched by the model.

We added support for multiple language models so that their probabilities, OOV penalties, and all other features are dynamically interpolated using MERT. This we use for the Haitian Creole-English tasks, where the first language model is a large model built on the monolingual data except SMS messages and the second small language model is built on the SMS messages. The OOV features play an important role here because frequent anonymization markers such as "[firstname]" do not appear in the large language model.

To scale to larger language models, we use BigFatLM[1], an open-source builder of large unpruned models with modified Kneser-Ney smoothing. Then, we filter the models to the system outputs. In order for an n-gram to be queried, all of the words must appear in system outputs for the same sentence. This enables a filtering constraint stronger than normal vocabulary filtering, which permits n-grams supported only by words in different sentences. Finally, we use KenLM (Heafield, 2011) for inference at runtime.

Our primary use of data is for language modeling. We used essientially every constrained resource available and appended them together to build one large model. For every language, we used the provided Europarl v6 (Koehn, 2005), News Crawl, and News Commentary corpora. In addition, we used:

English Gigaword Fourth Edition (Parker et al., 2009) and the English parts of United Nations documents, Giga-FrEn, and CzEng (Bojar and Žabokrtský, 2009) sections 0–7. For the Haitian Creole-English tasks, we built a separate language model on the SMS messages and used it alongside the large English model.

Czech CzEng (Bojar and Žabokrtský, 2009) sections 0–7

French Gigaword Second Edition (Mendonça et al., 2009a) and the French parts of Giga-FrEn and United Nations documents.

German There were no additional corpora available.

Spanish Gigaword Second Edition (Mendonça et al., 2009b) and the Spanish parts of United Nations documents.

4.2 Preprocessing

Many corpora contained excessive duplicate text. We wrote a deduplicator that removes all but the first instance of each line. Clean corpora generally reduced line count by 10-25% when deduplicated, resulting from naturally-occuring duplicates such as "yes ." We left the duplicate lines in these corpora. The News Crawl corpus showed a 72.6% reduction in line count due mainly to boilerplace, such as the

[1] https://github.com/jhclark/bigfatlm

Reuters comment section header and Fark headlines that appear in a box on many pages. We deduplicated the News Crawl corpus, United Nations documents, and New York Times and LA Times portions of English Gigaword.

The Giga-FrEn corpus is noisy. We removed lines from Giga-FrEn if any of the following conditions held:

- Invalid UTF8 or control characters.
- Less than 90% of characters are in the Latin alphabet (including diacritics) or punctuation. We did not count "<" and ">" as punctuation to limit the amount of HTML code.
- Less than half the characters are Latin letters.

System outputs and language model training data were normalized using the provided punctuation normalization script, Unicode codepoint collapsing, the provided Moses (Koehn et al., 2007) tokenizer, and several custom rules. These remove formatting-related tokens from Gigaword, rejoin some French words with internal apostrophes, and threshold repetitive punctuation. In addition, German words were segmented as explained in Section 4.3. Text normalization is more difficult for system combination because the system outputs, while theoretically detokenized, contain errors that result from different preprocessing at each site.

4.3 German Segmentation

German makes extensive use of compounding, creating words that do not cleanly align to English and have less reliable statistics. German-English translation systems therefore typically segment German compounds as a preprocessing step. In our case, we are concerned with combining translations into German that may be segmented differently. These can be due to stylistic choices; for example both "jahrzehnte lang" and "jahrzehntelang" appear with approximately equal frequency as shown in Table 1. Translation systems add additional biases due to the various preprocessing approaches taken by individual sites and inherent biases in models such as word alignment.

In order to properly align differently segmented words, we normalize by segmenting all system outputs and our language model training data using the single-best segmentation from cdec (Dyer et al., 2010). Running our system therefore produces segmented German output. Internally, we tuned towards segmented references but for final output it is desirable to rejoin compound words. Since the cdec segmentation was designed for German-English translation, no corresponding desegmenter was provided.

Words	Separate	Compounded
jahrzehnte lang	554	542
klar gemacht	840	802
unter anderem	49538	4
wieder herzustellen	513	1532

Table 1: Counts of separate or compounded versions of select words in the lowercased German monolingual data. Compounding can be optional or biased in either way.

We created a German desegmenter in the natural way: segment German words then invert the mapping to identify words that should be rejoined. To do so, we ran every word from the German monolingual data and system outputs through the cdec segmenter, counted both the compounded and segmented versions in the monolingual data, and removed those that appear segmented more often. Desegmenting is a mildly ambiguous process because n-grams to rejoin may overlap. When an n-gram compounded to one word, we gave that a score of n^2. The total score is a sum of these squares, favoring compounds that cover more words. Maximizing the score is a fast and exact dynamic programming algorithm. Casing of unchanged words comes from equally-weighted system votes at the character level while casing of rejoined words is based on the majority appearance in the corpus; this is almost always initial capital. We ran our desegmenter followed by the workshop's provided detokenizer to produce the submitted output.

5 Results

We tried many variations on the scheme, such as selecting different systems, tuning to BLEU (Papineni et al., 2002) or METEOR (Denkowski and Lavie, 2010), and changing the structure of the match count features from Section 2.3. To try these, we ran MERT 242 times, or about 24 times for each of the ten tasks in which we participated. Then we selected

the best performing systems on the tuning set and submitted them, with the secondary system chosen to meaningfully differ from the primary while still scoring well. Once the evaluation released references, we scored against them to generate Table 2.

On the featured Haitian Creole task, we show no and sometimes even negative improvement. This we attribute to the gap between the top system, bm-i2r, and the second place system. For htraw-en, where training data is noisy, the bm-i2r is 3.65 BLEU higher than the second place system at 28.53 BLEU. On htclean-en, the gap is 4.44 points to the second place cmu-denkowski-contrastive.

The main tasks were quite competitive and many systems were within a BLEU point of the top. This is an ideal scenario for system combination, and we show corresponding improvements. The English-Czech task is difficult for our scheme because we do not properly handle Czech morpology in alignment. On Czech-English, online-B beat other systems by a substantial (6.21 BLEU) margin, so we see little gain. On English-German, the gain is small but this is consistent with a general observation that more improvement is seen on higher-quality systems. Further, strength in this year's submission comes from language modeling, but only limited German data was available; segmenting German improved our scores. Translations into Spanish and French show the impact of Gigaword in those languages.

The evaluation's official metric is human ranking judgments. On this metric, our submissions score highest on eight of ten tracks: Czech-English, German-English, English-Czech, English-German, English-Spanish, English-French, the clean Haitian Creole-English task, and the raw Haitian Creole-English task. For Spanish-English, humans preferred RWTH's submission. For French-English, humans preferred RWTH and BBN. However, system combinations were ranked against other system combinations, but not against underlying systems, so we suspect that the bm-i2r submission still performs better than combinations on the Haitian Creole tasks. The human judges also preferred our translations more than BLEU (where we lead on three language pairs: English to German, Spanish, and French). We attribute this to the tendency of confusion networks to drop words supported by many systems due to position-based alignment er-

Track	Entry	BLEU	TER	MET
htraw-en	primary	32.30	56.57	61.05
	contrast	31.76	56.69	60.81
	bm-i2r	32.18	57.01	60.85
htclean-en	primary	36.39	51.16	63.72
	contrast	36.49	51.15	63.78
	bm-i2r	36.97	51.06	64.01
cz-en	primary	29.85	53.20	62.50
	contrast	29.88	53.19	62.40
	online-B	29.59	52.15	61.77
de-en	primary	26.21	56.19	60.56
	contrast	26.11	56.42	60.54
	online-B	24.30	57.95	59.63
es-en	primary	33.90	48.88	65.72
	contrast	33.47	49.41	66.41
	online-A	30.26	51.56	63.83
fr-en	primary	32.41	48.93	65.72
	contrast	32.15	49.12	65.71
	kit	30.36	50.74	64.32
en-cz	primary	20.80	61.17	41.68
	contrast	20.74	61.29	41.69
	online-B	20.37	61.38	41.40
en-de	primary	18.45	64.15	22.91
	contrast	18.27	64.48	22.75
	online-B	17.92	64.01	22.95
en-es	primary	36.47	47.08	34.96
	contrast	35.82	47.52	34.64
	online-B	33.85	50.09	33.96
en-fr	primary	36.42	48.28	24.29
	contrast	36.31	48.56	24.12
	online-B	35.34	48.68	23.53

Table 2: Automatic scores for our submissions. For comparison, the top individual system by BLEU is shown in the third row of each track. Test data and references were preprocessed prior to scoring. Metrics are uncased and METEOR 1.0 uses adequacy-fluency parameters. We show improvement on all tasks except Haitian Creole-English.

rors; our content-based alignment method avoids many of these errors. BLEU penalizes the missing word the same as missing punctuation while human judges will penalize heavily for missing content. For full results, we refer to the simultaneously published Workshop on Machine Translation findings paper.

6 Conclusion

We participated in the all ten tracks of the system combination, prioritizing participation and language support over optimizing for one particular language pair. Nonetheless, we show improvement on several tasks, including wins by BLEU on three tracks. The Haitian Creole and Czech-English tasks proved challenging due to the gap between top systems. However, other tracks show a variety of high-performing systems that make our scheme perform well. Unlike most other system combination schemes, our code is open source[2] so that these results may be replicated and brought to bear on similar problems.

Acknowledgements

Jon Clark assisted with language model construction and wrote BigFatLM. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 0750271 and by the DARPA GALE program.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings ACL*.

Ondřej Bojar and Zdeněk Žabokrtský. 2009. CzEng 0.9, building a large Czech-English automatic parallel treebank. *The Prague Bulletin of Mathematical Linguistics*, (92):63–83.

Michael Denkowski and Alon Lavie. 2010. Meteor-next and the meteor paraphrase tables: Improved evaluation support for five target languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 339–342, Uppsala, Sweden, July. Association for Computational Linguistics.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 7–12.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Kenneth Heafield and Alon Lavie. 2010. CMU multi-engine machine translation for WMT 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, UK, July. Association for Computational Linguistics.

Almut Silja Hildebrand and Stephan Vogel. 2009. CMU system combination for WMT'09. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 47–50, Athens, Greece, March. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, June.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*.

Ângelo Mendonça, David Graff, and Denise DiPersio. 2009a. French gigaword second edition. LDC2009T28.

Ângelo Mendonça, David Graff, and Denise DiPersio. 2009b. Spanish gigaword second edition. LDC2009T21.

Sushant Narsale. 2010. JHU system combination scheme for wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 311–314, Uppsala, Sweden, July. 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 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English gigaword fourth edition. LDC2009T13.

[2] `http://kheafield.com/code/mt`

Martin Porter. 2001. Snowball: A language for stemming algorithms. http://snowball.tartarus.org/.

Antti-Veikko Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2010. BBN system description for wmt10 system combination task. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 321–326, Uppsala, Sweden, July. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-2006)*, pages 223–231, Cambridge, MA, August.