

English → Czech System Combination



Ondřej Bojar

bojar@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics

Faculty of Mathematics and Physics

Charles University, Prague

- Overview of Evgeny's and Gregor's system combination.
- Motivation for our English→Czech datasets.
- Analysis of manual system combination.
- Main experiments:
 - Improving alignments.
 - Using Moses MERT.
 - Larger LMs and Tag LMs.
 - Small manual evaluation.
- Side tracks.
- Summary.

Rover System Combination (1/2)



Main idea of Fiscus (1997), extended by Matusov et al. (2008):
Systems vote which individual words should appear in the output.

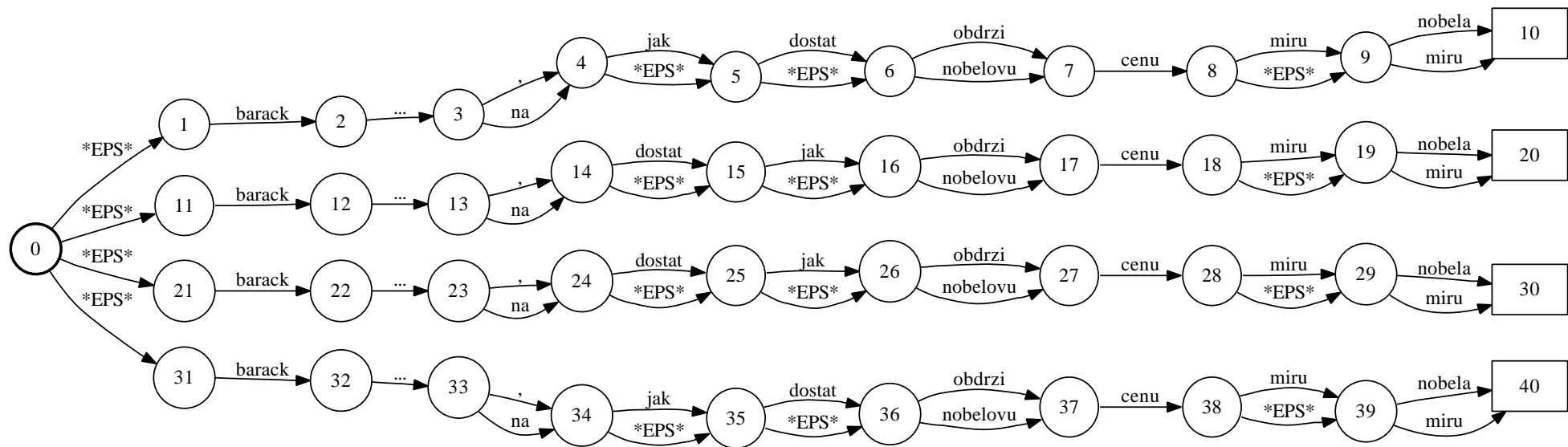
Procedure:

1. Given a “primary system” / “skeleton” ;
 - Align all systems to the skeleton (in bold), producing “bitexts” :
barack|**barack** . . . ,|**na** dostat| ϵ jak| ϵ nobelovu|**nobelovu** cenu|**cenu** míru|**míru**
barack|**barack** . . . na|**na** nobelovu|**nobelovu** cenu|**cenu** míru|**míru**
barack|**barack** . . . ,|**na** obdrží|**nobelovu** cenu|**cenu** míru| ϵ nobela|**míru**
 - Convert bitexts to confusion networks:

| | | | | | | | | |
|--------|-----|----|------------|------------|----------|------|------------|--------|
| barack | ... | na | ϵ | ϵ | nobelovu | cenu | ϵ | míru |
| <hr/> | | | | | | | | |
| barack | ... | , | dostat | jak | nobelovu | cenu | ϵ | míru |
| barack | ... | na | ϵ | ϵ | nobelovu | cenu | ϵ | míru |
| barack | ... | , | ϵ | ϵ | obdrží | cenu | míru | nobela |

Rover System Combination (2/2)

2. Combine confusion networks of various skeletons to one lattice:



3. Add language model scores.

4. Optimize weights (word penalty, LM, skeleton choice, . . .).

5. Select best path.

Combined Systems

In the following, we:

- Combine only ÚFAL's systems built for the WMT10 shared task.
- Tune and evaluate on WMT10 combination task datasets.

| | Dev Set | | Test Set | WMT10 Manual Rank |
|---------------|------------|---|------------|----------------------|
| bojar-primary | 16.00±1.15 | ↗ | 16.90±0.61 | 65.5 |
| bojar-sempos | 15.76±1.12 | ↗ | 16.61±0.59 | - |
| bojar-2step | 13.59±1.12 | ↗ | 14.38±0.58 | - |
| tectomt | 11.48±1.04 | ↗ | 13.19±0.58 | 60.1 |
| google | 17.32±1.25 | ↘ | 16.76±0.60 | 70.4 |
| eurotran | 9.64±0.92 | ↗ | 11.04±0.48 | 54.0 |
| pctrans2010 | 10.24±0.92 | ↗ | 10.84±0.46 | 62.1 |

Note Google discrepancy between Dev and Test \Rightarrow overfitting would be very likely.

“Bad” Systems Offer Words



Analyzing 44193 toks in the ref of WMT10 syscomb Test set.

- % tokens produced by bojar-primary?
- % tokens produced by one of the secondary systems only?

| | bojar-primary (16.90 ± 0.61) vs. | | | |
|----------------|--|------------------|------------------|-------------|
| | bojar-sempos | bojar-2stepsl | tectomt | the 3 other |
| | 16.61 ± 0.59 | 14.38 ± 0.58 | 13.19 ± 0.58 | - |
| In Both | 48.3 | 43.8 | 41.2 | 50.8 |
| Nowhere | 45.4 | 42.8 | 41.0 | 37.0 |
| Primary Only | 3.5 | 8.0 | 10.6 | 1.0 |
| Secondary Only | 2.8 | 5.4 | 7.1 | 11.2 |

- TectoMT could bring in up to 7.1% tokens, Two-Step 5.4% . . .
- The primary system alone has only 1.0% tokens on its own.
- Still 37% tokens of the reference not available.

To check the plausibility of “voting assumption” we manually do the task:

- Myself:
 - English→Czech, WMT10, 4 systems, 52 sents.
 - Reference translation available.
 - Attempted to stick to the original word order.
- Matusov (2009) (p. 140 talks about TC-STAR07 es→en):
 - Chinese(?)→English, IWSLT 2006, 4 systems, 489 sents.
 - Without looking at source or reference.
 - Allowed any reordering.
 - No further analysis beyond BLEU/TER/WER/PER.

Plausibility of Voting Assumption



How many produced tokens actually had the majority support?

| Supported by | Manual en→es | | Manual en→cs | | Auto en→cs | |
|--------------|-----------------|-------------|-----------------|-------------|---------------|-------------|
| | Toks | % | Toks | % | Toks | % |
| 1 | 978 | 15.8 | 160 | 19.4 | 30 | 3.6 |
| 2 | 1117 | 18.1 | 110 | 13.3 | 183 | 21.9 |
| ≤ 2 | 2095 | 33.9 | 270 | 32.7 | 213 | 25.5 |
| 3 | 1279 | 20.7 | 137 | 16.6 | 188 | 22.5 |
| 4 | 2806 | 45.4 | 417 | 50.6 | 435 | 52.0 |
| Total | 6180 | 100.0 | 824 | 100.0 | 836 | 100.0 |

... about $\frac{1}{3}$ of manually and $\frac{1}{4}$ of automatically combined tokens has no majority support (weights influence this).

Main Examined Directions



Baseline “system combination”:

- Add the 3 other outputs to training data of bojar-primary.

Within RWTH implementation (minor modifications):

- Improving word alignments.

RWTH alignment + Moses path selection and MERT:

- More detailed lattice arc weights.
- Handling of indicators in log-linear framework.
- Larger LMs.
- LMs for morphological tags.

Baseline Combinations



| Dataset | Test | Test | Dev |
|---------------------|------------|------------|------------|
| Weights | Default | Optimized | Default |
| Baseline RWTH | 17.50±0.64 | 17.42±0.63 | 16.28±1.20 |
| Add-to-training | - | 17.25±0.62 | 16.58±1.25 |
| Baseline RWTH+Moses | - | 17.19±0.61 | - |
| bojar-primary | - | 16.90±0.61 | 16.00±1.15 |
| google | - | 16.76±0.60 | 17.32±1.25 |

- RWTH marginally better unoptimized (sys. weights equal).
- MERT opt. in Moses worse than JaneOpt in RWTH setup.
Exceptionally, with milder pruning, Baseline RWTH+Moses got 17.57±0.61.
- Add-to-training works but very inefficient implementation:
 - Need to re-align, re-extract phrases, re-tune in MERT.

Improving Word Alignments



- GIZA++: No use of the fact that words are in the same lang.
- Using lemmas for Czech helps. (Bojar et al., 2006)

Baseline:

obdrží|nobelovu cenu|cenu míru| ϵ **nobela|míru**

Align lemmas and include an “equivalence dictionary”¹ in training:

obdrží|nobelovu cenu|cenu **míru|míru** nobela| ϵ

- Some misalignments fixed, some errors remained.
- Also tested automatically generated synonym classes.

¹E.g. *míru=míru* as a separate sentence.

Results of Improving Alignments

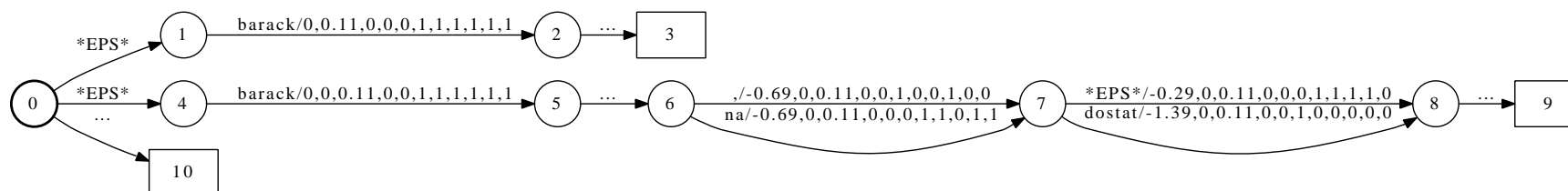


| | RWTH Optimizer | | Moses MERT | |
|-------------------|-------------------|-------------------|-------------------|-------------------|
| | Unoptimized | Optimized | Less Pruning | Dflt Pruning |
| Average±StdDev | 17.52±0.01 | 17.45±0.05 | 17.32±0.06 | 17.25±0.10 |
| eqvoc-lem-syndict | 17.52±0.63 | 17.51±0.62 | 17.30±0.60 | 17.16±0.60 |
| eqvoc-lem-syndict | 17.51±0.62 | 17.48±0.61 | 17.33±0.60 | 17.00±0.58 |
| eqvoc-lem-syndict | 17.52±0.63 | 17.48±0.62 | 17.21±0.60 | 17.29±0.59 |
| eqvoc-lem-syndict | 17.51±0.64 | 17.48±0.63 | 17.27±0.61 | 17.32±0.61 |
| eqvoc-stem3 | 17.52±0.63 | 17.48±0.62 | 17.41±0.64 | 17.35±0.62 |
| eqvoc-lem | 17.53±0.63 | 17.47±0.61 | 17.35±0.59 | 17.29±0.62 |
| eqvoc-lem-syndict | 17.53±0.63 | 17.47±0.62 | 17.26±0.61 | 17.29±0.60 |
| eqvoc-lem-syndict | 17.52±0.63 | 17.47±0.62 | 17.25±0.61 | 17.26±0.60 |
| eqvoc-stem4 | 17.52±0.63 | 17.47±0.62 | 17.36±0.61 | 17.07±0.60 |
| eqvoc-lem-syndict | 17.52±0.64 | 17.46±0.64 | 17.36±0.62 | 17.32±0.61 |
| eqvoc-lem-syndict | 17.51±0.63 | 17.46±0.63 | 17.26±0.61 | 17.33±0.60 |
| eqvoc-lem-syndict | 17.49±0.63 | 17.45±0.63 | 17.34±0.61 | 17.32±0.58 |
| lem | 17.50±0.63 | 17.45±0.63 | 17.27±0.60 | 17.37±0.61 |
| eqvoc | 17.51±0.64 | 17.44±0.63 | 17.27±0.59 | 17.18±0.59 |
| eqvoc-lem-syndict | 17.53±0.63 | 17.44±0.61 | 17.22±0.59 | 17.21±0.60 |
| eqvoc-lem-syndict | 17.53±0.63 | 17.44±0.63 | 17.37±0.61 | 17.33±0.60 |
| baseline | 17.50±0.64 | 17.42±0.63 | 17.57±0.61 | 17.19±0.61 |
| eqvoc-lem-syndict | 17.52±0.64 | 17.37±0.61 | 17.41±0.63 | 17.30±0.63 |

- Many variants of automatic synonym dict.
- Mixed results.
- Moses MERT less stable.

Lattice Arc Weights

- Current RWTH implementation has only 1 float per arc
⇒ scalar product done on the fly, hard to extend.
- Moses supports multiple weights and lattice input. (Dyer et al., 2008)
- I now create lattices myself, add several new weights:
 - Apriori-weight.** For each system and sentence (e.g. based on outside scores). So far not used.
 - Voting (RWTH).** The percentage of systems voting for this particular word at the given conf. net column
 - Sentence-level.** One for each system, indicating whether the system provided the skeleon. Collected incrementally along the sentence.
 - Arc-level.** One for each system, indicating how many output arcs were produced by the given system (incl. epsilon). These add up to voting-weight.
 - Primary-arcs.** How many output arcs are produced by the primary system
 - Primary-words (RWTH).** How many output words (i.e. arcs excl. eps.) are produced by the primary system.



Sentence-level flags have to be assigned per arc if we plan to determinize one day.

Indicators in Log-Linear Model

- Moses operates in log domain:
 - Scores added along the path and multiplied by weights.
 - Normalization: Divide each weight by $\sum |w_i|$.
- ⇒ The encoding of indicators influences search.

| | Probability Domain | | Log Domain | |
|------------|----------------------|-------------------|------------|--|
| | no | yes | no | yes |
| Bad | 0 | 1 | $-\infty$ | 0 |
| Common | $e^0 = 1$ | $e^1 \approx 2.7$ | 0 | 1 |
| Inverted | $e^1 \approx 2.7$ | $e^0 = 1$ | 1 | 0 <small>cf. tropical semiring</small> |
| Minus-Plus | $e^{-1} \approx 0.3$ | $e^1 \approx 2.7$ | -1 | 1 |

- Empirically Common/Inverted/Minus-Plus always differ but always fall within $\text{avg} \pm \text{stddev}$ ($3*7*18=378$ experiments).

Larger LMs

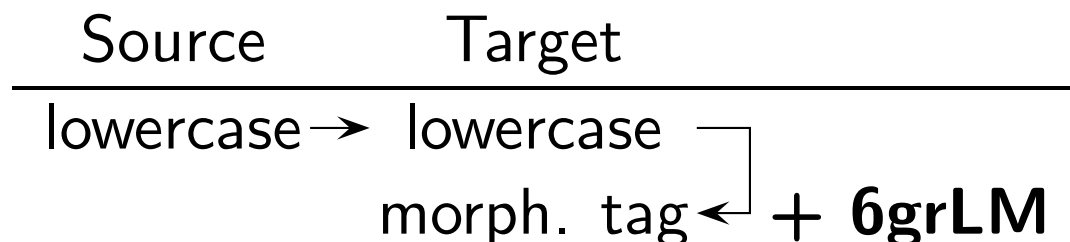
- By default, only 3gr LM based on combined hypotheses is used.
- RWTH saw no gains from using additional LM (G. Leusch, p.c.).
- en→cs and Moses MERT do make use of that.
- Additional data: WMT10mono, 13M sents, 211M tokens.

| | Baseline | Underlying Alignment Eqvoc+Lemmas | ⊗ ± σ Across All |
|---------------------|------------|--------------------------------------|------------------|
| RWTH Unoptimized | 17.50±0.64 | 17.53±0.63 | 17.52±0.01 |
| Moses +5grLM | 17.36±0.61 | 17.49±0.61 | 17.48±0.06 |
| Moses +4grLM | 17.63±0.59 | 17.45±0.62 | 17.46±0.08 |
| RWTH Optimized | 17.42±0.63 | 17.47±0.61 | 17.45±0.05 |
| Moses +3grLM | 17.46±0.61 | 17.44±0.63 | 17.41±0.07 |
| Moses Baseline | 17.32±0.63 | 17.34±0.61 | 17.32±0.06 |

- With the additional LM, Moses can reach RWTH optimizer.
- Higher n -grams marginally better.

LMs for Morphological Tags

- Bojar (2007) gains by using an additional LM over morphological tags in the factored translation (Koehn and Hoang, 2007).



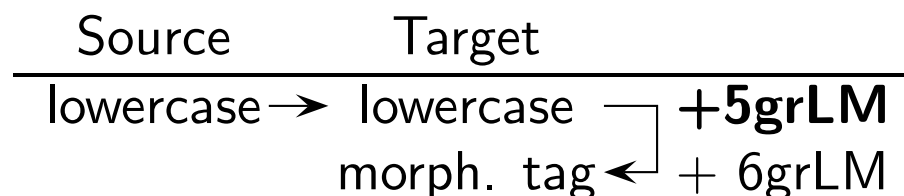
- Hypotheses are “tagged with unigram tagger” on the fly.

| | Underlying Alignment | | |
|---------------------------------|----------------------|-------------------|-------------------|
| | Baseline | Eqvoc+Lemmas | ⊗ ± σ Across All |
| Moses +tagLM, no Pruning | 17.88±0.62 | 17.95±0.59 | 17.90±0.12 |
| RWTH Unoptimized | 17.50±0.64 | 17.53±0.63 | 17.52±0.01 |
| RWTH Optimized | 17.42±0.63 | 17.47±0.61 | 17.45±0.05 |
| Moses Baseline | 17.32±0.63 | 17.34±0.61 | 17.32±0.06 |
| Moses +tagLM, with Pruning | 15.15±0.51 | - | - |

- Need to switch off beam pruning, tagged hyps wouldn't survive.

TagLM and Large LM

- We can combine TagLM and regular LM.
- This makes 15 weights in MERT optimization:
 - 9 arc weights, 3 LM weights, 2 tagger weights, word penalty.



| | Baseline | Underlying Alignment Eqvoc+Lemmas | ⊗ ± σ Across All |
|----------------------------|-------------------|--------------------------------------|-------------------|
| Moses +tagLM +5grLM | 18.01±0.66 | 17.80±0.59 | 17.97±0.09 |
| Moses +tagLM | 17.88±0.62 | 17.95±0.59 | 17.90±0.12 |
| RWTH Unoptimized | 17.50±0.64 | 17.53±0.63 | 17.52±0.01 |
| Moses +5grLM | 17.36±0.61 | 17.49±0.61 | 17.48±0.06 |
| RWTH Optimized | 17.42±0.63 | 17.47±0.61 | 17.45±0.05 |
| Moses Baseline | 17.32±0.63 | 17.34±0.61 | 17.32±0.06 |
| RWTH Optimized AllSys | 18.02±0.65 | 18.07±0.67 | - |

- In terms of BLEU score, this approaches the combination of all 7 systems.
- Incidentally, Moses +tagLM +5grLM using Minus-Plus got up to 18.26±0.64.

Manual Evaluation

- Manually ranked 65 sentences.
 - All the hyps get either one of equally-*, or
 - At least one hyp gets 1 and others get lower ranks.

| | | Equally | | Ranked as | | | |
|---------------------|-------------------|---------|----|-----------|----|----|---|
| | | Poor | Ok | 1 | 2 | 3 | 4 |
| Moses +tagLM +5grLM | 18.01±0.66 | 11 | 7 | 18 | 16 | 10 | 3 |
| RWTH Optimized | 17.42±0.63 | 11 | 7 | 22 | 17 | 7 | 1 |
| Moses Baseline | 17.32±0.63 | 11 | 7 | 17 | 14 | 14 | 2 |
| bojar-primary | 16.00±1.15 | 11 | 7 | 14 | 20 | 9 | 4 |

- Results unstable, would need many more sentences and annotators.
- Improved over single-best.

Training on Target-Side Data Only.

- See my abstract at MTMRL (Bojar and Tamchyna, 2011).
- Raised BLEU from 12.24 ± 0.44 to 12.65 ± 0.42 on a small dataset by training *TM* on target-side monolingual data.

Syntactic system combination. Idea by Carmen Heger.

- Use automatic CFGs and CFG-FSA intersection (Bar-Hillel et al., 1961) to score hyps by the grammar.

Use RWTH CRF tagger for my two-step translation.

- Thanks to Arne Mauser, experiments still run.

Jane for en→cs. Bad luck so far, very little time devoted.

Lessons Learned



Last time, I was praising Makefiles, SGE, . . .

The list is shorter this time, but still:

- Directory-local histories, `./history-bojar`
- SGE prologue and epilogue reporting usage.
- Inspired by your file caching tool to relieve NFS.
- OpenFST which I started using while here.

Btw, it's easy to expose tropical semiring over “power weights” in the command-line tools.

On the other hand:

- Scripting langs. are much more flexible than toolkits in C++.
- I'm happy there are ≤ 4 active cluster users in Prague.

Summary

- Learned to combine systems (voting over words).
... I would rather vote over “constituents”. \rightsquigarrow Future.
- Applied to en \rightarrow cs.
 - Moved to MERT optimization in Moses, more weights, LMs.
 - Improvement in BLEU thanks to TagLM.
 - Somewhat less convincing in manual evaluation.

Future:

- Will surely combine ÚFAL’s systems at next WMT.
- Hopefully with own implementation (align \rightarrow bitext is RWTH proprietary) or with e.g. Barrault (2010).

Again: Thanks for friendly and inspiring atmosphere.

References



Y. Bar-Hillel, M. Perles, and E. Shamir. 1961. On formal properties of simple phrase structure grammars. Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung, 14:143–172. Reprinted in Bar-Hillel's Language and Information - Selected Essays on their Theory and Application, Addison Wesley series in Logic, 1964, pp. 116-150.

Loic Barrault. 2010. MANY, Open Source Machine Translation System Combination. In Prague Bulletin of Mathematical Linguistics - Special Issue on Open Source Machine Translation Tools, number 93 in Prague Bulletin of Mathematical Linguistics. Charles University, January.

Ondřej Bojar and Aleš Tamchyna. 2011. Forms Wanted: Training SMT on Monolingual Data. Abstract at Machine Translation and Morphologically-Rich Languages. Research Workshop of the Israel Science Foundation University of Haifa, Israel, January.

Ondřej Bojar, Evgeny Matusov, and Hermann Ney. 2006. Czech-English Phrase-Based Machine Translation. In FinTAL 2006, volume LNAI 4139, pages 214–224, Turku, Finland, August. Springer.

Ondřej Bojar. 2007. English-to-Czech Factored Machine Translation. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 232–239, Prague, Czech Republic, June. Association for Computational Linguistics.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In Proceedings of ACL-08: HLT, pages 1012–1020, Columbus, Ohio, June. Association for Computational Linguistics.

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

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In Proc. of EMNLP.

References



Evgeny Matusov, Gregor Leusch, Rafael E. Banchs, Nicola Bertoldi, Daniel Dechelotte, Marcello Federico, Muntsin Kolss, Young-Suk Lee, Jose B. Marino, Matthias Paulik, Salim Roukos, Holger Schwenk, and Hermann Ney. 2008. System Combination for Machine Translation of Spoken and Written Language. IEEE Transactions on Audio, Speech and Language Processing, 16(7):1222–1237, September.

Evgeny Matusov. 2009. Combining Natural Language Processing Systems to Improve Machine Translation of Speech. Ph.D. thesis, RWTH Aachen University, Aachen, Germany, December.