Cutting the Long Tail: Hybrid Language Models for Translation Style Adaptation

Arianna Bisazza and Marcello Federico
Fondazione Bruno Kessler
Trento, Italy
{bisazza,federico}@fbk.eu

Abstract

In this paper, we address statistical machine translation of public conference talks. Modeling the style of this genre can be very challenging given the shortage of available in-domain training data. We investigate the use of a hybrid LM, where infrequent words are mapped into classes. Hybrid LMs are used to complement word-based LMs with statistics about the language style of the talks. Extensive experiments comparing different settings of the hybrid LM are reported on publicly available benchmarks based on TED talks, from Arabic to English and from English to French. The proposed models show to better exploit in-domain data than conventional word-based LMs for the target language modeling component of a phrase-based statistical machine translation system.

1 Introduction

The translation of TED conference talks is an emerging task in the statistical machine translation (SMT) community (Federico et al., 2011). The variety of topics covered by the speeches, as well as their specific language style, make this a very challenging problem.

Fixed expressions, colloquial terms, figures of speech and other phenomena recurrent in the talks should be properly modeled to produce translations that are not only fluent but that also employ the right register. In this paper, we propose a language modeling technique that leverages in-domain training data for style adaptation.

Hybrid class-based LMs are trained on text where only infrequent words are mapped to Part-of-Speech (POS) classes. In this way, topic-specific words are discarded and the model focuses on generic words that we assume more useful to characterize the language style. The factorization of similar expressions made possible by this mixed text representation yields a better n-gram coverage, but with a much higher discriminative power than POS-level LMs.

Hybrid LM also differs from POS-level LM in that it uses a word-to-class mapping to determine POS tags. Consequently, it doesn’t require the decoding overload of factored models nor the tagging of all parallel data used to build phrase tables. A hybrid LM trained on in-domain data can thus be easily added to an existing baseline system trained on large amounts of background data.

The proposed models are used in addition to standard word-based LMs, in the framework of log-linear phrase-based SMT.

The remainder of this paper is organized as follows. After discussing the language style adaptation problem, we will give an overview of relevant work. In the following sections we will describe in detail hybrid LM and its possible variants. Finally, we will present an empirical analysis of the proposed technique, including intrinsic evaluation and SMT experiments.

2 Background

Our working scenario is the translation of TED talks transcripts as proposed by the IWSLT Evaluation Campaign. This genre covers a variety of topics ranging from business to psychology. The available training material – both parallel and

1http://www.ted.com/talks

2http://www.iwslt2011.org
monolingual – consists of a rather small collection of TED talks plus a variety of large out-of-domain corpora, such as news stories and UN proceedings.

Given the diversity of topics, the in-domain data alone cannot ensure sufficient coverage to an SMT system. The addition of background data can certainly improve the n-gram coverage and thus the fluency of our translations, but it may also move our system towards an unsuitable language style, such as that of written news.

In our study, we focus on the subproblem of target language modeling and consider two English text collections, namely the in-domain TED and the out-of-domain NEWS\(^3\), summarized in Table 2. Because of its larger size – two orders of magnitude – the NEWS corpus can provide a better LM coverage than the TED on the test data. This is reflected both on perplexity and on the average length of the context (or history \(h\)) actually used by these two LMs to score the test’s reference translations. Note that the latter measure is bounded at the LM order minus one, and is inversely proportional to the number of back-offs performed by the model. Hence, we use this value to estimate how well an n-gram LM fits the test data. Indeed, despite the genre mismatch, the perplexity of a NEWS 5-gram LM on the TED-2010 test reference translations is 104 versus 112 for the in-domain LM, and the average history size is 2.5 versus 1.7 words.

| TED          | NEWS          |
|--------------|---------------|
| 1st [s] Thank you . [s] | 1st [s] Thank you . [s] |
| 2 [s] Thank you very much | 2 [s] Thank you very much |
| 3 [s] I’m going to | 3 [s] I’m going to |
| 4 [s] And I said , | 4 [s] And I said , |
| 5 [s] I don’t know | 5 [s] I don’t know |
| 6 [s] He said , “ | 6 [s] He said , “ |
| 7 [s] I said , “ | 7 [s] I said , “ |
| 8 [s] And of course , | 8 [s] And of course , |
| 9 [s] And one of the | 9 [s] And one of the |
| 10 [s] And I want to | 10 [s] And I want to |
| 11 [s] And that’s what | 11 [s] And that’s what |
| 12 [s] We’re going to | 12 [s] We’re going to |
| 13 [s] And I think that | 13 [s] And I think that |
| 14 [s] And you can see | 14 [s] And you can see |
| 15 [s] And this is a | 15 [s] And this is a |
| 16 [s] And this is the | 16 [s] And this is the |
| 17 [s] And he said , | 17 [s] And he said , |
| 18 [s] So this is a | 18 [s] So this is a |

Table 3: Excerpts from TED and NEWS training vocabularies, as ranked by frequency. Numbers denote the frequency rank.

Yet we observe that the style of public speeches is much better represented in the in-domain corpus than in the out-of-domain one. For instance, let us consider the vocabulary distribution\(^4\) of the

\(^3\)http://www.statmt.org/wmt11/translation-task.html

\(^4\)Hesitations and filler words, typical of spoken language, are not covered in our study because they are generally not reported in the TED talk transcripts.
two corpora (Table 3). The very first forms, as ranked by frequency, are quite similar in the two corpora. However, there are important exceptions: the pronouns I and you are among the top 20 frequent forms in the TED, while in the NEWS they are ranked only 40th and 64th respectively. Other interesting cases are the words actually, stuff, guy and amazing, all ranked about 10 times higher in the TED than in the NEWS corpus.

We can also analyze the most typical ways to start and end a sentence in the two text collections. As shown in Table 1, the frequency ranking of sentence-initial and sentence-final 5-grams in the in-domain corpus is notably different from the out-of-domain one. TED’s most frequent sentence-initial 5-gram “[s] Thank you . /s]” is not at all attested in the NEWS corpus. As for the 4th most common sentence start “[s] And I said . ,” is only ranked 8199th in the NEWS, and so on. Notably, the top ranked NEWS 5-grams include names of cities (Washington, New York) and of news agency (AP, Reuters). As regards sentence endings, we observe similar contrasts: for instance, the word sequence “and so on . /s]” is ranked 4th in the TED and 157th in the NEWS while “., you know . /s]” is 5th in the TED and only 1652th in the NEWS.

These figures confirm that the talks have a specific language style, remarkably different from that of the written news genre. In summary, talks are characterized by a massive use of first and second persons, by shorter sentences, and by more colloquial lexical and syntactic constructions.

3 Related Work

The brittleness of n-gram LMs in case of mismatch between training and task data is a well known issue (Rosenfeld, 2000). So called domain adaptation methods (Bellegarda, 2004) can improve the situation, once a limited amount of task specific data become available. Ideally, domain-adaptive LMs aim to improve model robustness under changing conditions, involving possible variations in vocabulary, syntax, content, and style. Most of the known LM adaption techniques (Bellegarda, 2004), however, address all these variations in a holistic way. A possible reason for this is that LM adaptation methods were originally developed under the automatic speech recognition framework, which typically assumes the presence of one single LM. The progressive adoption of the log-linear modeling framework in many NLP tasks has recently introduced the use of multiple LM components (features), which permit to naturally factor out and integrate different aspects of language into one model. In SMT, the factored model (Koehn and Hoang, 2007), for instance, permits to better tailor the LM to the task syntax, by complementing word-based n-grams with a part-of-speech (POS) LM, that can be estimated even on a limited amount of task-specific data. Besides many works addressing holistic LM domain adaptation for SMT, e.g. Foster and Kuhn (2007), recently methods were also proposed to explicitly adapt the LM to the discourse topic of a talk (Ruiz and Federico, 2011). Our work makes another step in this direction by investigating hybrid LMs that try to explicitly represent the speaking style of the talk genre. As a difference from standard class-based LMs (Brown et al., 1992) or the more recent local LMs (Monz, 2011), which are used to predict sequences of classes or word-class pairs, our hybrid LM is devised to predict sequences of classes interleaved by words. While we do not claim any technical novelty in the model itself, to our knowledge a deep investigation of hybrid LMs for the sake of style adaptation is definitely new. Finally, the term hybrid LM was inspired by Yazgan and Saraclar (2004), which called with this name a LM predicting sequences of words and sub-words units, devised to let a speech recognizer detect out-of-vocabulary-words.

4 Hybrid Language Model

Hybrid LMs are n-gram models trained on a mixed text representation where each word is either mapped to a class or left as is. This choice is made according to a measure of word commonness and is univocal for each word type.

The rationale is to discard topic-specific words, while preserving those words that best characterize the language style (note that word frequency is computed on the in-domain corpus only). Mapping non-frequent terms to classes naturally leads to a shorter tail in the frequency distribution, as visualized by Figure 1. A model trained on such data has a better n-gram coverage of the test set and may take advantage of a larger context when scoring translation hypotheses.

As classes, we use deterministically assigned POS tags, obtained by first tagging the data with
This is actually a 20

And this is actually 12

And that’s actually 10

NN, it’s actually 9

we’re actually going to 8

Table 4: Most common hybrid 5-grams containing the words guy and actually, along with absolute frequency.

4.4 Handling morphology

Token frequency-based measures may not be suitable for languages other than English. When translating into French, for instance, we have to deal with a much richer morphology.

As a solution we can use lemmas, univocally assigned to word types in the same manner as POS tags. Lemmas can be employed in two ways: only for word selection, as a frequency measure, or also for word representation, as a mapping for common words. In the former, we preserve reflected variants that may be useful to model the language style, but we also risk to see n-gram coverage decrease due to the presence of rare types. In the latter, only canonical forms and POS tags

5This differs from the tf-idf widely used in information retrieval, which is used to measure the relevance of a term in a document. Instead, we measure commonness of a term in the whole corpus.
appear in the processed text, thus introducing a further level of abstraction from the original text.

Here follows a TED sentence in its original version (first line) and after three different hybrid mappings – namely $W_P=.25$, $W_P=.25$ with lemma forms, and $W_P=.50$:

| Hybrid | $|V|$ | POS-Err | $\bar{H}_{10g}$ |
|--------|------|--------|--------------|
| all words | 51299 | 0.0% | 1.7 |
| all lemmas | 38486 | 0.0% | 1.9 |
| .25 POS/words | 475 | 1.9% | 2.7 |
| .50 POS/words | 93 | 4.1% | 3.5 |
| .75 POS/words | 50 | 5.7% | 4.1 |
| allPOS | 43 | 6.6% | 4.4 |
| .25 POS/lemmas | 302 | 1.8% | 2.8 |
| .25 POS/words(idf) | 301 | 1.9% | 2.7 |

Table 5: Comparison of LMs obtained from different hybrid mappings of the English TED corpus: vocabulary size, POS error rate, and average word history on IWSLT-tes2010’s reference translations.

5 Evaluation

In this section we perform an intrinsic evaluation of the proposed LM technique, then we measure its impact on translation quality when integrated into a state-of-the-art phrase-based SMT system.

5.1 Intrinsic evaluation

We analyze here a set of hybrid LMs trained on the English TED corpus by varying the ratio of POS-mapped words and the word representation technique (word vs lemma). All models were trained with the IRSTLM toolkit (Federico et al., 2008), using a very high n-gram order (10) and Witten-Bell smoothing.

First, we estimate an upper bound of the POS tagging errors introduced by deterministic tagging. At this end, the hybrily mapped data is compared with the actual output of Tree Tagger on the TED training corpus (see Table 5). Naturally, the impact of tagging errors correlates with the ratio of POS-mapped tokens, as no error is counted on non-mapped tokens. For instance, we note that the POS error rate is only 1.9% in our primary setting, $W_P=.25$ and word representation, whereas on a fully POS-mapped text it is 6.6%. Note that the English tag set used by Tree Tagger includes 43 classes.

Now we focus on the main goal of hybrid text representation, namely increasing the coverage of the in-domain LM on the test data. Here too, we measure coverage by the average length of word history $\bar{H}$ used to score the test reference translations (see Section 2). We do not provide perplexity figures, since these are not directly comparable across models with different vocabularies. As shown by Table 5, n-gram coverage increases with the ratio of POS-mapped tokens, ranging from 1.7 on an all-words LM to 4.4 on an all-POS LM. Of course, the more words are mapped, the less discriminative our model will be. Thus, choosing the best hybrid mapping means finding the best trade-off between coverage and informativeness.

We also applied hybrid LM to the French language, again using Tree Tagger to create the POS mapping. The tag set in this case comprises 34 classes and the POS error rate with $W_P=.25$ is 1.2% (compare with 1.9% in English). As previously discussed, morphology has a notable effect on the modeling of French. In fact, the vocabulary reduction obtained by mapping all the words to their most probable lemma is -45% (57959 to 31908 types in the TED corpus), while in English it is only -25%.

5.2 SMT baseline

Our SMT experiments address the translation of TED talks from Arabic to English and from English to French. The training and test datasets were provided by the organizers of the IWSLT11 evaluation, and are summarized in Table 6. Marked in bold are the corpora used for hybrid LM training. Dev and test sets have a single reference translation.

For both language pairs, we set up competitive phrase-based systems using the Moses toolkit (Koehn et al., 2007). The decoder features a statistical log-linear model including a phrase translation model and a phrase reordering model (Tillmann, 2004; Koehn et al., 2005), two word-based language models, distortion, word and phrase penalties. The translation and reordering models are obtained by combining models independently trained on the available parallel.
| Corpus        | |S| | |W| | 7 |
|---------------|-------------------|-------------------|-------------------|
| AR-EN TED     | 90K               | 1.7M              | 18.9              |
| EN TED        | 124K              | 2.4M              | 19.5              |
| EN NEWS       | 30.7M             | 782M              | 25.4              |
| AR test       | dev2010           | 934               | 19K               |
|               | tst2010           | 1664              | 30K               |
| EN-FR TED     | 105K              | 2.0M              | 19.5              |
| EN-FR NEWS    | 11K               | 291M              | 26.5              |
| FR TED        | 107K              | 2.2M              | 20.6              |
| FR NEWS       | 11.6M             | 291M              | 25.2              |
| EN test       | dev2010           | 934               | 20K               |
|               | tst2010           | 1664              | 32K               |

Table 6: IWSLT11 training and test data statistics: number of sentences |S|, number of tokens |W| and average sentence length 7. Token numbers are computed on the target language, except for the test sets.

The Arabic Treebank tokenization scheme isolates conjunctions w+ and /+, prepositions f+, k+, b+, future marker s+, pronominal suffixes, but not the article Al+.

translation models, while the English-French system uses lowercased models and a standard re-casing post-process.

Feature weights are tuned on dev2010 by means of a minimum error training procedure (MERT) (Och, 2003). Following suggestions by Clark et al. (2011) and Cettolo et al. (2011) on controlling optimizer instability, we run MERT four times on the same configuration and use the average of the resulting weights to evaluate translation performance.

5.3 Hybrid LM integration

As previously stated, hybrid LMs are trained only on in-domain data and are added to the log-linear decoder as an additional target LM. To this end, we use the class-based LM implementation provided in Moses and IRSTLM, which applies the word-to-class mapping to translation hypotheses before LM querying. The order of the additional LM is set to 10 in the Arabic-English evaluation and 7 in the English-French, as these appeared to be the best settings in preliminary tests.

Translation quality is measured by BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006). To test whether differences among systems are statistically significant we use approximate randomization as done in (Riezler and Maxwell, 2005).

Model variants. The effect on MT quality of various hybrid LM variants is shown in Table 7. Note that allPOS and allLemmas refer to deterministically assigned POS tags and lemmas, respectively. Concerning the ratio of POS-mapped tokens, the best performing values are Wp=.25 in Arabic-English and Wp=.50 in English-French. These hybrid mappings outperform all the uniform representations (words, lemmas and POS) with statistically significant BLEU and METEOR improvements.

The fdif experiment involves the use of document frequency for the selection of common words. Its performance is very close to that of hy-

8Detailed instructions on how to build and use hybrid LMs can be found at http://hlt.fbk.eu/people/bisazza.

9We use case-sensitive BLEU and TER, but case-insensitive METEOR to enable the use of paraphrase tables distributed with the tool (version 1.3).

10Translation scores and significance tests were computed with the Multeval toolkit (Clark et al., 2011): https://github.com/jhclark/multeval.
brid LMs simply based on term frequency; only METEOR gains 0.1 points in Arabic-English. A possible reason for this is that document frequency was computed on fixed-size text chunks rather than on real document boundaries (see Section 4.1). The lemmaF experiment refers to the use of canonical forms for frequency measuring; this technique does not seem to help in either language pair. Finally, we compare the use of lemmas versus surface forms to represent common words. As expected, lemmas appear to be helpful for French language modeling. Interestingly this is also the case for English, even if by a small margin (+0.2 METEOR, -0.1 TER).

Summing up, hybrid mapping appears as a winning strategy compared to uniform mapping. Although differences among LM variants are small, the best model in Arabic-English is .25-POS/lemmas, which can be thought of as a domain-generic lemma-level LM. In English-French, instead, the highest scores are achieved by .50-POS/lemmas or .50-POS/lemmas(fdf), that is POS-level LM with few frequently occurring lexical anchors (vocabulary size 59). An interpretation of this result is that, for French, modeling the syntax is more helpful than modeling the style. We also suspect that the French TED corpus is more irregular and diverse with respect to the style, than its English counterpart. In fact, while the English corpus include transcripts of talks given by English speakers, the French one is mostly a collection of (human) translations. Typical features of the speech style may have been lost in this process.

Comparison with baseline. In Table 8 the best performing hybrid LM is compared against the baseline that only includes the standard LMs described in Section 5.2. To complete our evaluation, we also report the effect of an in-domain LM trained on 50 word classes induced from the corpus by maximum-likelihood based clustering (Och, 1999).

In the two language pairs, both types of LM result in consistent improvements over the baseline. However, the gains achieved by the hybrid approach are larger and all statistically significant. The hybrid approach is significantly better than the unsupervised one by TER in Arabic-English and by BLEU and METEOR in English-French (these significances are not reported in

| Added InDomain 10gLM | BLEU ↑ | MET ↑ | TER ↓ |
|----------------------|--------|--------|--------|
| .00 POS/words (all words)† | 26.1 | 30.5 | 55.4 |
| .00 POS/lemmas (all lem.)† | 26.0 | 30.5 | 55.4 |
| 1.0 POS/words (all POS)† | 25.9 | 30.6 | 55.3 |
| .25 POS/words† | 26.5 | 30.6 | 54.7 |
| .50 POS/words | 26.5 | 30.6 | 54.9 |
| .75 POS/words | 26.3 | 30.7 | 55.0 |
| .25 POS/words(fdf) | 26.5 | 30.7 | 54.7 |
| .25 POS/lemmasF | 26.4 | 30.6 | 54.8 |
| .25 POS/lemmas | 26.5 | 30.8 | 54.6 |

Table 7: Comparison of various hybrid LM variants. Translation quality is measured with BLEU, METEOR and TER (all in percentage form). The settings used for weight tuning are marked with †. Best models according to all metrics are highlighted in bold.

| Added InDomain 7gLM | BLEU ↑ | MET ↑ | TER ↓ |
|---------------------|--------|--------|--------|
| .00 POS/words (all words)† | 31.1 | 52.5 | 49.9 |
| .00 POS/lemmas (all lem.)† | 31.2 | 52.6 | 49.7 |
| 1.0 POS/words (all POS)† | 31.4 | 52.8 | 49.8 |
| .25 POS/lemmas† | 31.5 | 52.9 | 49.7 |
| .50 POS/lemmas | 31.9 | 53.3 | 49.5 |
| .75 POS/lemmas | 31.7 | 53.2 | 49.6 |
| .50 POS/lemmas(fdf) | 31.9 | 53.3 | 49.5 |
| .50 POS/lemmasF | 31.6 | 53.0 | 49.6 |
| .50 POS/words | 31.7 | 53.1 | 49.5 |

Table 8: Final MT results: baseline vs unsupervised word classes-based LM and best hybrid LM. Statistically significant improvements over the baseline are marked with • at the p < .01 and ◦ at the p < .05 level.
the table for clarity). The proposed method appears to better leverage the available in-domain data, achieving improvements according to all metrics: +0.5/+0.4/-1.0 BLEU/METEOR/TER in Arabic-English and +0.7/-0.6/-0.3 in English-French, without requiring any bitext annotation or decoder modification.

**Talk-level analysis.** To conclude the study, we analyze the effect of our best hybrid LM on Arabic-English translation quality, at the single talk level. The test used in the experiments (tst2010) consists of 11 transcripts with an average length of 151±73 sentences. For each talk, we compare the baseline BLEU score with that obtained by adding a .25-POS/lemmas hybrid LM. Results are presented in Figure 2. The dark and light columns denote baseline and hybrid-LM BLEU scores, respectively, and refer to the left y-axis. Additional data points, plotted on the right y-axis in reverse order, represent talk-level perplexities (PP) of a standard 5-gram LM trained on TED (○) and those of the .25-POS/lemmas 10-gram hybrid LM (△), computed on reference translations.

What emerges first is a dramatic variation of performance among the speeches, with baseline BLEU scores ranging from 33.95 on talk “00” to only 12.42 on talk “02”. The latter talk appears as a corner case also according to perplexities (397 by word LM and 111 by hybrid LM). Notably, the perplexities of the two LMs correlate well with each other, but the hybrid’s PP is much more stable across talks: its standard deviation is only 14 points, while that of the word-based PP is 79. The BLEU improvement given by hybrid LM, however modest, is consistent across the talks, with only two outliers: a drop of -0.2 on talk “00”, and a drop of -0.7 on talk “02”. The largest gain (+1.1) is observed on talk “10”, from 16.8 to 17.9 BLEU.

### 6 Conclusions

We have proposed a language modeling technique that leverages the in-domain data for SMT style adaptation. Trained to predict mixed sequences of POS classes and frequent words, hybrid LMs are devised to capture typical lexical and syntactic constructions that characterize the style of speech transcripts.

Compared to standard language models, hybrid LMs generalize better to the test data and partially compensate for the disproportion between in-domain and out-of-domain training data. At the same time, hybrid LMs show more discriminative power than merely POS-level LMs. The integration of hybrid LMs into a competitive phrase-based SMT system is straightforward and leads to consistent improvements on the TED task, according to three different translation quality metrics.

Target language modeling is only one aspect of the statistical translation problem. Now that the usability of the proposed method has been assessed for language modeling, future work will address the extension of the idea to the modeling of phrase translation and reordering.

### Acknowledgments

This work was supported by the T4ME network of excellence (IST-249119), funded by the DG INFSO of the European Commission through the 7th Framework Programme. We thank the anonymous reviewers for their valuable suggestions.

### References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan, June. 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 the 40th Annual Meeting of the Association of Computational Linguistics (ACL), pages 311–318, Philadelphia, PA.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 57–64, Ann Arbor, Michigan, June. Association for Computational Linguistics.

R. Rosenfeld. 2000. Two decades of statistical language modeling: where do we go from here? Proceedings of the IEEE, 88(8):1270–1278.

Nick Ruiz and Marcello Federico. 2011. Topic adaptation for lecture translation through bilingual latent semantic models. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 294–302, Edinburgh, Scotland, July. Association for Computational Linguistics.

Nick Ruiz, Arianna Bisazza, Fabio Brugnara, Daniele Falavigna, Diego Giuliani, Suhel Jaber, Roberto Gretter, and Marcello Federico. 2011. FBK @ IWSLT 2011. In International Workshop on Spoken Language Translation (IWSLT), San Francisco, CA.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In Proceedings of International Conference on New Methods in Language Processing.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In 5th Conference of the Association for Machine Translation in the Americas (AMTA), Boston, Massachusetts, August.

Christoph Tillmann. 2004. A Unigram Orientation Model for Statistical Machine Translation. In Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL).

A. Yazgan and M. Saraçlar. 2004. Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition. In Proceedings of ICASSP, volume 1, pages 1–745–8 vol.1, may.