Unsupervised Alignment for Segmental-based Language Understanding

Stéphane Huet and **Fabrice Lefèvre**
Université d'Avignon, LIA-CERI, France
{stephane.huet,fabrice.lefevre}@univ-avignon.fr

Abstract

Recent years' most efficient approaches for language understanding are statistical. These approaches benefit from a segmental semantic annotation of corpora. To reduce the production cost of such corpora, this paper proposes a method that is able to match first identified concepts with word sequences in an unsupervised way. This method based on automatic alignment is used by an understanding system based on conditional random fields and is evaluated on a spoken dialogue task using either manual or automatic transcripts.

1 Introduction

One of the very first step to build a spoken language understanding (SLU) module for dialogue systems is the extraction of literal concepts from word sequences hypothesised by a speech recogniser. To address this issue of concept tagging, several techniques are available. These techniques rely on models, now classic, that can be either discriminant or generative. Among these, we can cite: hidden Markov models, finite state transducers, maximal entropy Markov models, support vector machines, dynamic Bayesian networks (DBNs) or conditional Markov random fields (CRFs) (Lafferty et al., 2001). In (Hahn et al., 2011), it is shown that CRFs obtain the best performance on a reference task (MEDIA) in French (Bonneau-Maynard et al., 2005), but also on two other comparable corpora in Italian and Polish. Besides, the comparison of the understanding results of manually vs automatically transcribed utterances has shown the robustness of CRFs.

Among the approaches evaluated in (Hahn et al., 2011) was a method using log-linear models comparable to those used in stochastic machine translation, which turned out to have lower performance than CRF. In this paper, we further exploit the idea of applying automatic translation techniques to language understanding but limiting ourselves to the objective of obtaining a segmental annotation of training data.

In many former approaches literal interpretation was limited to list lexical-concept relations; for instance this is the case of the PHOENIX system (Ward, 1991) based on the detection of keywords. The segmental approach allows a finer-grained analysis considering sentences as segment sequences during interpretation. This characteristic enables the approach to correctly connect the various levels of sentence analysis (lexical, syntactic and semantic). However, in order to simplify its practical application, segments have been designed specifically for semantic annotation and do not integrate any constraint in their relation with the syntactic units (chunks, phrasal groups, etc.). Not only it simplifies the annotation process itself but as the overall objective is to use the interpretation module inside a spoken dialogue system, transcribed speech data are noisy and generally bound the performance of syntactic analysers (due to highly spontaneous and ungrammatical utterances from the users, combined with errors from the speech recognizer).

Among other interesting proprieties, segmental approaches offer a convenient way to dissociate the detection of a conceptual unit from the estimation of its associated value. The value corresponds to the normalisation of the surface form. For instance, if

Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, pages 97–104,
Edinburgh, Scotland, UK, July 27–31, 2011.

the segment "no later than eleven" is associated with the concept `departure-time`, its value is "morning"; the same value is associated with the segments "between 8 and noon" or "in the morning". The value estimation requires a link between concepts and sentence words. Then it becomes possible to treat the normalisation problem by means of regular expressions or concept-dependent language models (allowing an integrated approach such as described in (Lefèvre, 2007)). In the case of global approaches (not segmental), value detection must be directly incorporated in the conceptual units to identify, as in (Mairesse et al., 2009). The additional level is a real burden and is only affordable when the number of authorised values is low.

Obviously a major drawback of the approach is its cost: associating concept tags with a dialogue transcription is already a tedious task and its complexity is largely increased by the requirement for a precise delimitation of the support (lexical segment) corresponding to each tag. The SLU evaluation campaign MEDIA has been the first opportunity to collect and distribute a reasonably-sized corpus endowed with segmental annotations.

Anyhow the difficulty remains unchanged each time a corpus has to be collected for a new task. We propose in this study a new method that reduces the effort required to build training data for segmental annotation models. Making the assumption that the concepts evoked in a sentence are automatically detected beforehand or provided by an expert, we study how to associate them with their lexical supports without *prior* knowledge. A conceptual segmental annotation is obtained using alignment techniques designed to align multilingual parallel corpora in the machine translation domain. This annotation can be considered as unsupervised since it is done without a training corpus with links between word sequences and concepts.

We present in the paper the necessary adaptations for the application of the alignment techniques in this new context. They have been kept to their minimal so as to maintain the highest level of generality, which in return benefits from the availability of existing software tools. Using a reference annotation, we evaluate the alignment quality from the unsupervised approach in two interesting situations depending on whether the correct order of the concepts is known or not. Finally, the end-to-end evaluation of the approach is made by measuring the impact of the alignments on the CRF-based understanding system.

After a brief recall of the conceptual decoding principles in Section 2, the principles of automatic alignment of parallel corpora are described in Section 3 along with the specificities due to the alignment of semantic concepts. Section 4 presents the experiments and comments on the results, while Section 5 concludes the paper.

2 Segmental conceptual decoding

If literal interpretation can be seen as the translation of natural language to the set of semantic tag sequences, then the methods and models of machine translation can be used. Since the number of concepts is generally much lower than the vocabulary size, this particular type of translation can also be considered as a mere classification problem in which the conceptual constituents represent the class to identify. Interpretation can thus be performed by methods and models of classification.

Discriminant approaches model the conditional probability distribution of the semantic constituent sequence (or concepts) $c_1 \dots c_n$ considering a word sequence $w_1 \dots w_T$: $P(c_1^n|w_1^T)$. In generative approaches, the joint probability $P(c_1^n, w_1^T)$ is modelized instead and can be used to compute inferences either for prediction/decoding or parameter training.

Generative models (such as hidden Markov models) have been first introduced to address the understanding problem with stochastic approaches (Levin and Pieraccini, 1995). Recent variants offer more degrees of freedom in modeling (see for instance (He and Young, 2005) or (Lefèvre, 2007)). Since then log-linear models have clearly shown their superiority for tasks of sequence tagging (Hahn et al., 2011).

Several variants of log-linear models differ in their conditional variable independence assumptions and use different normalisation steps. CRFs (Lafferty et al., 2001) represent linear chains of random independent variables, all conditioned over the entire sequence and the normalisation is global over the sequence.

Some generative approaches such as DBNs make inferences in multi-level models (Lefèvre, 2007)

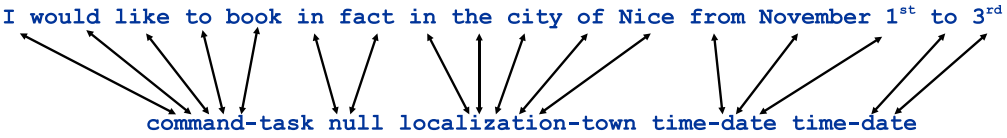

Figure 1: Example of an alignment of words with their conceptual units.

and intrinsically take into account segmentation. For models unable to handle multi-level representations (as CRF), it is convenient to represent segments directly at the tag level. For this purpose the BIO formalism can be used: B is added to tags starting a segment, I to tags inside a segment and O to out-of-domain tags (if these are not already handled through a specific NULL tag). In the case displayed in Figure 1, the concept sequence becomes: `B-cmd-task I-cmd-task I-cmd-task B-null I-null B-loc-town I-loc-town I-loc-town I-loc-town I-loc-town B-time-date I-time-date B-time-date I-time-date I-time-date`.

3 Semantic concept alignment

Automatic alignment is a major issue in machine translation. For example, word-based alignments are used to generate phrase tables that are core components for many current statistical machine translation systems (Koehn et al., 2007). The alignment task aims at finding the mapping between words of two sentences in relation of translation. It faces several difficulties:

- some source words are not associated with a translated word;
- others are translated by several words;
- matched words may occur at different positions in both sentences according to the syntactic rules of the considered languages.

Several statistical models have been proposed to align two sentences (Brown et al., 1993). One of their main interests is their ability to be built in an unsupervised way from a parallel corpus aligned at the sentence level, but not at the word level. Formally, from a sentence $S = s_1 \dots s_m$ expressed in a source language and its translation $T = t_1 \dots t_n$ expressed in a target language, an IBM-style alignment $A = a_1 \dots a_m$ connects each source word to a target word ($a_j \in \{1, ..., n\}$) or to the so-called NULL token which accounts for untranslated target words. IBM statistical models evaluate the translation of S into T from the computation of $P(S, A|T)$; the best alignment $\hat{A}$ can be deduced from this criterion using the Viterbi algorithm:

$$\hat{A} = \text{argmax}_A P(S, A|T) \ . \quad (1)$$

IBM models differ according to their complexity level. IBM1 model makes the strong assumption that alignments are independent and can be evaluated only through the transfer probabilities $P(s_i|t_j)$. The HMM model, which is an improvement over IBM2, adds a new parameter $P(a_j|a_{j-1}, n)$ that assumes a first-order dependency between alignment variables. The next models (IBM3 to IBM5) are mainly based on two types of parameters:

- distortion, which measures how words of T are reordered with respect to the index of the words from S they are aligned with,
- fertility, which measures the usual number of words that are aligned with a target word t_j.

In order to improve alignments, IBM models are usually applied in both translation directions. These two alignments are then symmetrized by combining them. This last step is done via heuristic methods; a common approach is to start with the intersection and then iteratively add links from the union (Och et al., 1999).

If we have at our disposal a method that can find concepts contained in an utterance, segmental annotation can be obtained by aligning words $S = w_1^T$ with the found concepts $T = c_1^n$ (Fig. 1). Concepts are ideally generated in the correct order with respect to the word segments of the analysed utterance. In a more pragmatic way, concepts are likely to be produced as bag-of-concepts rather than ordered sequences.

Statistical alignment methods used in machine translation are relevant in our context if we consider that the target language is the concept language. There are nevertheless differences with genuine language translation. First, each word is aligned to at most one concept, while a concept is aligned with one word or more. Consequently, it is expected that word fertilities are one for the alignment of words toward concepts and concept fertilities are one or more in the reverse direction. Another consequence is that NULL words are useless in our context. These specificities of the alignment process raise some difficulties with regard to IBM models. Indeed, according to the way probabilities are computed, the alignment of concepts toward words only allows one word to be chosen per concept, which prevents this direction from having a sufficient number of links between words and concepts.

Another significant difference with translation is related to the translated token order. While word order is not random in a natural language and follows syntactic rules, it is not the case anymore when a word sequence have to be aligned with a bag-of-concepts. HMM and IBM2 to IBM5 models have parameters that assume that the index of a matched source word or the indices of the translations of the adjacent target words bear on the index of target words. Therefore, the randomness of the concept indices can disrupt performance obtained with these models, contrary to IBM1. As shown in the next section, it is appropriate to find ways to explicitly re-order concept sequences than to let the distortion parameters handle the problem alone.

4 Experiments and results

4.1 Experimental setup

The evaluation of the introduced methods was carried out on the MEDIA corpus (Bonneau Maynard et al., 2008). This corpus consists of human-machine dialogues collected with a wizard of Oz procedure in the domain of negotiation of tourist services. Produced for a realistic task, it is annotated with 145 semantic concepts and their values (more than 2k in total for the enumerable cases). The audio data are distributed with their manual transcripts and automatic speech recognition (ASR) hypotheses. The corpus is divided into three parts: a training set (approximatively 12k utterances), a development set (1.2k) and a test set (3k).

The experiments led on the alignment methods were evaluated on the development corpus using MGIZA++ (Gao and Vogel, 2008), a multi-thread version of GIZA++ (Och and Ney, 2003) which also allows previously trained IBM alignments models to be applied on the development and test corpora.[1] The conceptual tagging process was evaluated on the test corpus, using WAPITI (Lavergne et al., 2010) to train the CRF models. Several setups have been tested:

- manual vs ASR transcriptions,
- inclusion (or not) of values during the error computation.

Several concept orderings (before automatic alignment) have also been considered:

- a first ideal one, which takes reference concept sequences as they are, aka **sequential order**;
- two more realistic variants that sort concepts either **alphabetically** or **randomly**, in order to simulate bag-of-concepts. Alphabetical order is introduced solely to show that a particular order (which is not related to the natural order) might misled the alignment process by introducing undue regularities.

To give a rough idea, these experiments required a few minutes of computing time to train alignment models of 12k utterances, a few hours to train CRF models (using 8 CPUs on our cluster of Xeon CPUs) and a few seconds to apply alignment and CRF models in order to decode the test corpus.

4.2 Experimental results for alignment

Alignment quality is estimated using the *alignment error rate* (AER), a metric often employed in machine translation (Och and Ney, 2000). If H stands for hypothesis alignments and R for reference alignments, AER is computed by the following relation:[2]

$$AER = 1 - \frac{2 \times |H \cap R|}{|H| + |R|} \quad . \qquad (2)$$

[1] With `previousa`, `previoust`, `previousn`, etc parameters.

[2] This equation is a simplification of the usually provided one because all alignments are considered as sure in our case.

In our context, this metrics is evaluated by representing a link between source and target identities by (w_i, c_j), instead of the usual indices (i, j). Indeed, alignments are then used to tag words. Besides, concepts to align have positions that differ from the ones in the reference when they are reordered to simulate bags-of-concepts.

As mentioned in the introduction, we resort to widely used tools for alignment in order to be as general as possible in our approach. We do not modify the algorithms and rely on their generality to deal with specificities of the studied domain. To train iteratively the alignment models, we use the same pipeline as in MOSES, a widely used machine translation system (Koehn et al., 2007):

1. 5 iterations of IBM1,
2. 5 iterations of HMM,
3. 3 iterations of IBM3 then
4. 3 iterations of IBM4.

To measure the quality of the built models, the model obtained at the last iteration of this chain is applied on the development corpus.

All the words of an utterance should normally be associated with one concept, which makes the IBM models' NULL word useless. However, in the MEDIA corpus, a `null` semantic concept is associated with words that do not correspond to a concept relevant for the tourist domain and may be omitted by counting on the probability with the NULL word included in the IBM models. Two versions were specifically created to test this hypothesis: one with all the reference concept sequences and another without the `null` tags. The results measured when taking into account these tags (AER of 14.2 %) are far better than the ones obtained when they are discarded (AER of 27.4 %), in the word → concept alignment direction.[3] We decided therefore to keep the `null` in all the experiments.

Table 1 presents the alignment results measured on the development corpus according to the way concepts are reordered with respect to the reference and according to the considered alignment direction. The three first lines exhibit the results obtained with the last IBM4 iteration. As expected, the AER measured with this model in the concept → word direction (second line), which can only associate at most one word per concept, is clearly higher than the one obtained in the opposite direction (first line). Quite surprisingly, an improvement in terms of AER (third line) over the best direction (first line) is observed using the default MOSES heuristics (called *grow-diag-final*) that symmetrizes alignments obtained in both directions.

IBM1 models, contrary to other models, do not take into account word index inside source and target sentences, which makes them relevant to deal with bag-of-concepts. Therefore, we measured how AER varies when using models previously built in the training chain. The results obtained by applying IBM1 and by symmetrizing alignments (last line), show finally that these simple models lead to lower performance than the one measured with IBM4 or even HMM (last line), the concepts being ordered alphabetically or randomly (two last columns).

The previous experiments have shown that alignment is clearly of lower quality when algorithms are faced with bags-of-concepts instead of well-ordered sequences. In order to reduce this phenomenon, sequences are reordered after a first alignment $\mathcal{A}_1$ generated by the symmetrized IBM4 model. Two strategies have been considered to fix the new position of each concept c_i. The first one averages the indices of the words w_i that are aligned with c_i according to $\mathcal{A}_1$:

$$\text{pos}_1(c_j) = \frac{\sum_{is.t.(i,j)\in\mathcal{A}_1} i}{\text{Card}(\{(i,j) \in \mathcal{A}_1\})} \ . \qquad (3)$$

The second one weights each word index with their transfer probabilities determined by IBM4:

$$\text{pos}_2(c_j) = \frac{\sum_{is.t.(i,j)\in\mathcal{A}_1} i \times f(w_i, c_j)}{\sum_{is.t.(i,j)\in\mathcal{A}_1} f(w_i, c_j)} \qquad (4)$$

where

$$f(w_i, c_j) = \lambda P(c_j|w_i) + (1-\lambda)P(w_i|c_j) \qquad (5)$$

and λ is a coefficient fixed on the development corpus.

Training alignment models on the corpus reordered according to pos_1 (Tab. 2, second column)

[3] For a fair comparison between both setups, the `null` concept was ignored in H and R for this series of experiments.

	Sequential order	Alphabetic order	Random order
word → concept IBM4	14.4	29.2	28.6
concept → word IBM4	40.9	51.6	49.0
symmetrized IBM4	**12.8**	**27.3**	**25.7**
symmetrized IBM1	33.2	33.2	33.1
symmetrized HMM	14.8	29.9	28.7

Table 1: AER (%) measured on the MEDIA development corpus with respect to the alignment model used and its direction.

	Initial	1st reordering iteration		Last reordering iteration
		pos_1	pos_2	pos_2
Alphabetic order	27.3	22.2	21.0	**19.4**
Random order	25.7	21.9	20.2	**18.5**

Table 2: AER (%) measured on the MEDIA development corpus according to the strategy used to reorder concepts.

or pos_2 (third column) leads to a significant improvement of the AER. This reordering step can be repeated as long as performance goes on improving. By proceeding like this until step 3 for the alphabetic order and until step 7 for the random order, values of AER below 20 % (last column) are finally obtained. It is noteworthy that random reordering has better results than alphabetic reordering. Indeed, HMM, IBM3 and IBM4 models have probabilities that are more biased in this latter case, where the same sequences occur more often although many are not in the reference.

4.3 Experimental results for spoken language understanding

In order to measure how spoken language understanding is disturbed by erroneous alignments, CRFs parameters are trained under two conditions: one where concept tagging is performed by an expert and one where corpora are obtained using automatic alignment. The performance criterion used to evaluate the understanding task is the *concept error rate* (CER). CER is computed in a similar way as word error rate (WER) used in speech recognition; it is obtained from the Levenshtein alignment between both hypothesized and reference sequences as the ratio of the sum of the concepts in the hypothesis substituted, inserted or omitted on the total number of concepts in the manual reference annotation. The `null` concept is not considered during the score computation. The CER can also take into account the normalized values in addition to the concept tags.

Starting from a state-of-the-art system (Manual column), degradations due to various alignment conditions are reported in Table 3. It can be noted that the absolute increase in CER is at most 8.0 % (from 17.6 to 25.6 with values) when models are trained on the corpus aligned with IBM models; the ordering information brings it back to 3.7 % (17.6 to 21.3), and finally with automatic transcription the impact of the automatic alignments is smaller (resp. 5.8 % and 2.0 %). As expected random order is preferable to alphabetic order (slight gain of 1 %).

In Table 4, the random order alignments are used but this time the n-best lists of alignments are considered and not only the 1-best hypotheses. Instead of training CRFs with only one version of the alignment for a concept-word sequence pair, we filter out from the n-best lists the alignments having a probability above a given threshold. It can be observed that varying this confidence threshold allows an improvement of the SLU performance (CER can be reduced by 0.8 % for manual transcription and 0.4 % for automatic transcription). However, this improvement is not propagated to scores with values (CER was reduced at best by 0.1 for manual transcription and was increased for automatic tran-

	Manual	Automatic alignments		
		Sequential	Alphabetic order	Random order
Manual transcription	13.9 (17.6)	17.7 (21.3)	22.6 (26.4)	22.0 (25.6)
ASR transcription (wer 31 %)	24.7 (29.8)	27.1 (31.8)	31.5 (36.4)	30.6 (35.6)

Table 3: CER (%) measured for concept decoding on the MEDIA test corpus with several alignment methods of the training data. Inside parenthesis, CER for concepts and values.

scription). After closer inspection of the scoring alignments, an explanation for this setback is that the manually-designed rules used for value extraction are perturbed by loose segmentation. This is particularly the case for the concept used to annotate co-references, which has confusions between the values `singular` and `plural` (e.g. "this" is singular and "those" plural). This issue can be solved by an *ad hoc* adaptation of the rules. However, it would infringe our objective of relying upon unsupervised approaches and minimizing human expertise. Therefore, a better answer would be to resort to a probabilistic scheme also for value extraction (as proposed in (Lefèvre, 2007)).

The optimal configuration (confidence threshold of 0.3, 4th row of Table 4) is close to the baseline 1-best system in terms of the number of training utterances. We also tried a slightly different setup which adds the filtered alignments to the former corpus before CRF parameter training (i.e. the 1-best is not filtered in the n-best list). In that case performance remains pretty stable with respect to the filtering process (CER is around 21.4 % for concepts and 25.2 % for concept+value for thresholds between 0.1 and 0.7).

5 Conclusion

In this study an unsupervised approach is proposed to the problem of conceptual unit alignment for spoken language understanding. We show that unsupervised statistical word alignment from the machine translation domain can be used in this context to associate semantic concepts with word sequences. The quality of the derived alignment, already good in the general case (< 20 % of errors on the word-concept associations), is improved by knowledge of the correct unit order (< 15 %). The impact of automatic alignments on the understanding performance is an absolute increase of +8 % in terms of CER, but is reduced to less than +4 % in the ordered case. When automatic transcripts are used, these gaps decrease to +6 % and below +3 % respectively. From these results we do believe that the cost vs performance ratio is in favour of the proposed method.

Acknowledgements

This work is partially supported by the ANR funded project PORT-MEDIA.[4]

References

Hélène Bonneau-Maynard, Sophie Rosset, Christelle Ayache, Anne Kuhn, and Djamel Mostefa. 2005. Semantic annotation of the MEDIA corpus for spoken dialog. In *Proceedings of Eurospeech*, pages 3457–3460, Lisboa, Portugal.

Hélène Bonneau Maynard, Alexandre Denis, Frédéric Béchet, Laurence Devillers, F. Lefèvre, Matthieu Quignard, Sophie Rosset, and Jeanne Villaneau. 2008. MEDIA : évaluation de la compréhension dans les systèmes de dialogue. In *L'évaluation des technologies de traitement de la langue, les campagnes Technolangue*, pages 209–232. Hermès, Lavoisier.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, OH, USA.

Stefan Hahn, Marco Dinarelli, Christian Raymond, Fabrice Lefèvre, Patrick Lehen, Renato De Mori, Alessandro Moschitti, Hermann Ney, and Giuseppe Riccardi. 2011. Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, 19(6):1569–1583.

[4] `www.port-media.org`

	# train utterances	Manual transcription	ASR transcription (WER = 31 %)
1-best	12795	22.0 (25.6)	30.6 (**35.6**)
filtered 10-best (conf thres = 0.1)	18955	21.7 (25.8)	31.2 (36.9)
filtered 10-best (conf thres = 0.2)	15322	21.3 (**25.5**)	30.7 (36.3)
filtered 10-best (conf thres = 0.3)	13374	**21.2** (25.7)	**30.2** (36.0)
filtered 10-best (conf thres = 0.5)	10963	21.4 (25.7)	30.6 (36.2)
filtered 10-best (conf thres = 0.7)	9647	25.4 (29.1)	32.9 (38.2)

Table 4: CER (%) measured for concept decoding on the MEDIA test corpus with filtered n-best lists of random order alignments of the training data. Inside parenthesis, CER for concepts and values.

Yulan He and Steve Young. 2005. Spoken language understanding using the hidden vector state model. *Speech Communication*, 48(3–4):262–275.

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 *Proceedings of ACL, Companion Volume*, pages 177–180, Prague, Czech Republic.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pages 282–289, Williamstown, MA, USA.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical very large scale CRFs. In *Proceedings of ACL*, pages 504–513, Uppsala, Sweden.

Fabrice Lefèvre. 2007. Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation. In *Proceedings of ICASSP*, Honolulu, Hawai.

Esther Levin and Roberto Pieraccini. 1995. Concept-based spontaneous speech understanding system. In *Proceedings of Eurospeech*, pages 555–558, Madrid, Spain.

François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2009. Spoken language understanding from unaligned data using discriminative classification models. In *Proceedings of ICASSP*, Taipei, Taiwan.

Franz Joseph Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of Coling*, volume 2, pages 1086–1090, Saarbrücken, Germany.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, College Park, MD, USA.

Wayne Ward. 1991. Understanding spontaneous speech: the Phoenix system. In *Proceedings of ICASSP*, pages 365–368, Toronto, Canada.