# Confidence measures in dialogue annotation by N-gram transducers

**Carlos-D. Martínez-Hinarejos, Vicent Tamarit, José-Miguel Benedí**

Instituto Tecnológico de Informática, Universitat Politècnica de València, Camino de Vera s/n, 46022, Valencia, Spain
{cmartine,vtamarit,jbenedi}@dsic.upv.es

**Abstract**

Dialogue annotation is a necessary step for the development of dialogue systems, specially for data-based dialogue strategies. Manual annotation is hard and time-consuming, and automatic techniques can be used to obtain a draft annotation and speed-up the process. An interesting feature in this framework is the presentation of the draft annotation with confidence levels on the correctness of every part of the hypothesis can make even faster the supervision process. In this paper we propose a method to calculate confidence measures for an automatic dialogue annotation model, and test it for the annotation of a task-oriented human-computer corpus on railway information. The results show that our proposal is a good starting point for incorporating confidence measures in the dialogue annotation process.

**Keywords:** Dialogue annotation, confidence measures, N-gram transducers

## 1. Introduction

In the natural language processing field, a dialogue system is defined as a computer system that interacts with a human being by using dialogue (Lee et al., 2010). Most dialogue systems employ only speech, and are used in many applications such as information systems that are accessed by telephone (Seneff and Polifroni, 2000) (e.g., ticket reservation systems, timetable consultation systems, etc.).

These systems require a model of the dialogue structure in order to mimic this structure in the automatic system. Therefore, a framework for the dialogue structure must be defined. One of the most popular frameworks is that based on the speech act theory (Austin, 1962), that focuses on the communicative acts performed in the dialogue interaction. From this framework, the concept of dialogue act (DA) (Bunt, 1994) is derived; a DA is a label which codes the intention of the current interaction, along with its corresponding data related to the task. Since in each interaction several intentions can be distinguished in different subsequences, each of these subsequences (called segments) has an associated DA. In an automatic dialogue systems, DA labels can be associated to both computer and human user.

The annotation of a set of dialogues in terms of DA is an important task for the obtainment of data-based automatic dialogue systems, since these systems are based on statistical models which learn the relation between the dialogue state and the DA labels (Williams and Young, 2007). Many annotation schemes have been proposed in several projects, such as DASML (Core and Allen, 1997) or DATE (Walker and Passonneau, 2001). In any case, manual annotation of dialogue corpora by human experts is required, but this is a hard and long task. Thus, in the last years some automatic techniques have been proposed to obtain a draft annotation and speed up the annotation process. The current most promising technique is based on the N-gram Transducers (NGT) model (Tamarit et al., 2011).

However, all these automatic techniques are not error-free, and a human supervision of the annotation must be performed. In this process, it will be very helpful to warn to the human expert about those parts in the proposed hypothesis that, according to the automatic technique, are more error-prone (i.e., the technique has somehow an evidence that its proposal may not be correct). Thus, the reviewer can concentrate on those parts that are probably wrong and avoids to spend time in reviewing parts likely to be correct. The usual tool for deciding which part is likely to be correct or not are the confidence measures.

Confidence measures have been very popular in Automatic Speech Recognition (ASR) (Jiang, 2005) (where they are applied on recognised words), and in the last few years they have been extended into other NLP fields such as Machine Translation (Ueffing et al., 2003) or parsing (Sánchez-Sáez et al., 2009). In dialogue systems, the use of confidence measures has been mainly directed to the use of ASR confidence measures to improve the reaction of the system on recognition errors and misunderstandings (San-Segundo et al., 2001), but as far as we know no clear application of confidence measures was proposed for automatic dialogue annotation techniques, and more specifically for NGT.

In this work we propose a statistically-based formulation of confidence measures for dialogue annotation, and we implement and evaluate these proposals in the NGT technique. In Section 2., an overview of the NGT model is provided. In Section 3., the formulation of the confidence measures is presented. In Section 4., the experimental corpus is detailed. In Section 5., experiments are described and results are showed and analysed. In Section 6., conclusions and future work lines are described.

## 2. The N-gram Transducers model for dialogue annotation

The annotation of a dialogue transcription can be formulated as an optimisation problem: given a word sequence $\mathcal{W}$ that represents a dialogue, the aim is to obtain the sequence of DA labels $\mathcal{U}$ that maximises the posterior probability $\Pr(\mathcal{U}|\mathcal{W})$. Since dialogue transcriptions are usually presented in turns, if a dialogue has $T$ turns, we can express the dialogue as the sequence of words of the turns, i.e., $\mathcal{W} = W_1^T = W_1 W_2 \cdots W_T$; the same decomposition can be applied to DA sequences, i.e., $\mathcal{U} = U_1^T = U_1 U_2 \cdots U_T$. In these sequences, $W_t$ and $U_t$ represent the sequence of words and DA labels, respectively, for turn $t$ of the dialogue. $W_t^s$ and $U_t^s$ will represent the sequence of words and DA labels between turns $t$ and $s$, both included.

As a result, the optimisation problem can be expressed as:

$$\widehat{\mathcal{U}} = \operatorname*{argmax}_{\mathcal{U}} \Pr(\mathcal{U}|\mathcal{W}) = \operatorname*{argmax}_{\mathcal{U}_1^T} \Pr(\mathcal{U}_1^T|\mathcal{W}_1^T) \qquad (1)$$

Among several options, this problem can be decomposed using the Bayes' rule (as presented in (Martínez-Hinarejos et al., 2008)) or directly solved by other models such as the N-gram Transducers (NGT) model (Tamarit et al., 2011). The NGT model employs a n-gram model that acts as a transducer; the n-gram model is obtained by following the inference process defined by the GIATI [1] Stochastic Finite-State Transducers (SFST) inference technique (Casacuberta et al., 2005).

GIATI forms, from a corpus of input-output aligned training sentences, an extended training corpus formed by the combination of the input and output words (this process is known as re-labelling); from this extended corpus, a smoothed n-gram model is inferred. This n-gram model can be converted into a SFST by undoing the re-labelling process (see details in (Casacuberta et al., 2005)), but the NGT technique proposes its direct use as a transducer in its n-gram form. This avoids the difficulties of modelling the smoothing probabilities in a finite-state model, and it is easy to apply when no cross-inverted alignments are present in the original training corpus of aligned sentences.

The search process for NGT is a Viterbi process in which, apart from the NGT model itself, a n-gram model for the output language is included. The search process forms a search tree, where the $i$-th level is associated to the $i$-th input word, and each input word is expanded into as many children nodes as different output had associated in the training process. For example, if a word $w$ was associated to outputs $o_1$ and $o_2$, apart from the empty output, when $w_i = w$ in the $i$-th level each node of the tree will produce three children nodes (one for the empty output, another for $o_1$, and another for $o_2$).

The probability of each branch in the search process is updated according to the probability of the parent node, the probability of the NGT model and the probability of the output n-gram model. This last probability is taken into account only when the child presents an output. At the end of the search process, the node of the final level with the highest probability is chosen and its associated branch is retrieved, which gives a sequence of extended words that provide both the output symbols and a segmentation of the input sequence.

The NGT model can be applied to dialogue annotation by using as input language the words of the transcribed dialogue, and as output language the corresponding DA labels. Input and output are converted into the extended corpus by attaching the DA label to the last word of the corresponding segment (using a metasymbol such as @). The results obtained by the NGT in dialogue annotation are outstanding with respect to those obtained with more classical techniques (such as Hidden Markov Models (Stolcke et al., 2000)). More details on the NGT model and the search process can be consulted in (Martínez-Hinarejos et al., 2009; Tamarit et al., 2011).

## 3. Confidence measures

Although the NGT model provides good results in dialogue annotation, for a practical interactive system of dialogue annotation (in which a human annotator provides the final correct annotation) it is important to provide a guide to the user on how confident is the automatic annotation system with respect to its hypothesis. This is similar to what happens in ASR systems which are used in speech transcription, where each word in the decoding result can be signalled to improve the performance of the correctness by the human transcriptor. These guidelines given by the system are based on the so-called *confidence measures*: a score between 0 and 1 which evaluates how confident is the system in a segment of the decoded hypothesis.

Our proposal on confidence measures is based on the result of the Viterbi decoding process. After the decoding process of a dialogue with $l$ words, the best path (the branch with highest probability) can be seen as a sequence of extended words $e_1 \cdot e_2 \cdots e_l$, where each extended word $e_i$ is given as the local solution of decoding input word $w_i$. In general, we can state the problem as obtaining the confidence of the output $E = e_i^j$ that occurs when processing the subsequence $w_i^j$ of the input sequence $w$. Since we are usually interested in the confidence of each extended word, the problem is usually reduced to $E = e_i$ (the extended word given for $w_i$) and $j = i + 1$. We will denote the event of producing $E$ between times $i$ and $j$ as $C_{ij}^E$.

Following a probabilistic approximation, the confidence of the event $C_{ij}^E$ for the input sequence $w$ can be taken as the posterior probability of the occurrence of that event given $w$, i.e., $\Pr(C_{ij}^E|w)$ (Wessel et al., 2001). In that case, using the Bayes' rule:

$$\Pr(C_{ij}^E|w) = \frac{\Pr(C_{ij}^E, w)}{\Pr(w)} \qquad (2)$$

The joint probability in the numerator can be expressed in the terms of the Forward-Backward computation (Devijver, 1985) with the corresponding $\alpha$ (forward) and $\beta$ (backward) terms, thus giving:

$$\Pr(C_{ij}^E|w) = \frac{\alpha_i(E) f(E,i,j) \beta_j(E)}{\Pr(w)} \qquad (3)$$

Each term has the following meaning:

- $\alpha_i(E) = \Pr(w_1, \ldots, w_{i-1}, q_i = E|\lambda)$; that is, the probability that, given the model $\lambda$, the sequence $w_1, \ldots, w_{i-1}$ is processed and the state $q_i$ is reached and produces the output $E$.
- $f(E,i,j) = \Pr(w_i, \ldots, w_{j-1}|\lambda)$; that is, the probability of processing $w_i, \ldots, w_{j-1}$ given the model.

[1] Grammatical Inference and Alignment for Transducer Infer

- $\beta_j(E) = \Pr(w_{j-1}, \ldots, w_l | q_j = E, \lambda)$; that is, the probability of, given the model $\lambda$ and that the output $E$ reached the state $q_j$, the sequence $w_{j-1}, \ldots, w_l$ is processed.
- $\Pr(w) = \Pr(w_1, \ldots, w_l)$; that is, the probability of the input word sequence according to the model.

In the terms of the NGT model, all these terms can be computed as it follows:

- $\alpha_i(E)$ is the sum of all the probabilities of the nodes at level $i$ that produce output $E$.
- $f(E, i, j)$ is the sum of all the probabilities of the transitions between all nodes at level $i$ and all nodes at level $j$ where the output $E$ was produced.
- $\beta_j(E)$ is the sum of all the $\beta$ computations of the children nodes of all nodes at level $j$ that were reached by producing the output $E$.
- $\Pr(w)$ is the sum of the probabilities of all the different solutions that the search process produced for the input sequence $w$ (that is, the sum of the probabilities of all the nodes of the last level).

With these definitions, Equation 3 can be used to obtain the confidence measure of each output. Equation 3 presents a normalisation factor (all the terms are divided by $\Pr(w)$). Thus, these original computations must be properly normalised to obtain the correct values of the confidence. We will refer to this measure as *Forward-Backward confidence measure* (FBCM).

Another possibility is to assume that only the local decision on the current level is good enough to properly compute a confidence measure. This approximation requires less computation, since the forward and backward terms are ignored, and can be used to obtain a faster value of the confidence on the current hypothesis. In this case, the forward and backward terms can be neglected from the formulation of the confidence measure, thus giving:

$$\Pr(C_{ij}^E | w) = \frac{f(E, i, j)}{\Pr(w)} \qquad (4)$$

The term $f(E, i, j)$ is computed in the same manner as for the FBCM, with the proper normalisation given by $\Pr(w)$ in Equation 4. We will refer to this measure as *Transition-based confidence measure* (TransCM).

## 4. Experimental data

The calculation of confidence measures for dialogue annotation was performed on the Dihana corpus (Benedí et al., 2006). Dihana is a dialogue corpus composed of 900 task-oriented human-computer telephone dialogues in Spanish. The corpus is oriented to obtaining information about long-distance railway services in Spain, which covers items such as timetables, fares and additional services for the trains. This corpus was acquired using the Wizard of Oz (WoZ) technique (Fraser and Gilbert, 1991), in which a human expert simulates the behaviour of an automatic system. There were a total of 225 voluntary speakers, which performed the acquisition without restrictions; the only semantic restriction was provided by the scenario they had to accomplish, which varied from acquisition to acquisition. The acquisition process resulted in 6,280 user turns and 9,133 system turns, with a vocabulary of about 900 words and a total of 5.5 hours of speech signal. The dialogues were manually transcribed and annotated with DA at the segment level by using an annotation scheme that is presented in (Alcácer et al., 2005). This scheme defines each DA as a combination of three different levels (speech act, concept and argument, initially defined in (Fukada et al., 1998). The mean number of segments (label ocurrences) per turn is about 1.5, and a total number of 248 labels (153 for user turns and 95 for system turns) were defined. When only the first two levels are considered, the number of labels reduces to 72 labels (45 for user and 27 for system).

## 5. Experiments and results

A series of experiments was defined in order to examine the performance of the proposed confidence measures in the annotation by the NGT technique of the Dihana corpus. The confidence measures were implemented in the current version of the NGT software [2], and this new software was applied to the Dihana corpus.

The Dihana corpus was preprocessed to reduce its complexity using a process similar to that reported in previous works (Martínez-Hinarejos et al., 2009): all the words were transcribed to lowercase, a categorisation (which included times, dates, town names, fares, etc.) was performed, the words were speaker-labelled (U for user, S for system), and punctuation marks were separated from words. A cross-validation approach was followed by defining 5 partitions of 180 dialogues each partition. The annotation models that were employed were a 4-gram for the NGT model and a 3-gram of DA as output language model, since the best results reported for Dihana with the NGT model use these models (Martínez-Hinarejos et al., 2009). Since in the training samples many input words have only associated a possible output (including the empty output), these words must be excluded in the evaluation of the confidence measure (since they always will have a total confidence). For all the other words, the possible outputs depend on the training set, but they are a subset of the extended words that can be formed by the word itself and the word attached to any of the DA labels (248 for the 3-level labelling and 72 for the 2-level labelling).

An initial measure of the quality of the annotation technique can be obtained by computing the Classification Error Rate (CER) measure for the set without single-output words in the baseline experiment (i.e., where no confidence measures are used and every decision is taken a confident enough). In the CER computation the possible events are:

- No output is in the hypothesis:
  - No output is in the reference: correct.
  - Output is in the reference: incorrect.

[2] Available in `http://www.dsic.upv.es/~cmartine/research/resources.html`.

Table 1: Dihana CER results (2 and 3-level labels) for NGT model with a 4-gram for NGT and a 3-gram for DA language model.

| 2 levels | 3 levels |
|---|---|
| 3.81 | 8.48 |

Table 2: AROC measure for the FBCM and TransCM confidence measures, for the 2 and 3 levels label set of Dihana. Baseline is 50. The NGT model was a 3-gram and the DA output language model was a 4-gram.

| Confidence measure | 2 levels | 3 levels |
|---|---|---|
| FBCM | 90.59 | 83.91 |
| TransCM | 91.16 | 83.89 |

- An output is in the hypothesis:
  - No output is in the reference: incorrect.
  - Output in the reference, but different: incorrect.
  - Output in the reference, and the same: correct.

CER is expressed as the percentage of incorrect events with respect to the total number of events. The corresponding results are presented in Table 1. This CER is quite low, which suggest that even the use of high-quality confidence measures will not produce improvements in this measure. However, the use of CER for the evaluation of confidence measures is unfaithful in many cases, and consequently we will use other wide-accepted evaluation measures that rely on the concepts of correct and incorrect events as well.

The evaluation was performed by using the classical "Receiver Operating Characteristic" (ROC) curves (Egan, 1975) and the "Area under ROC curve" (AROC). A ROC curve is a measure which represents the *true rejection rate* (the proportion of the truly incorrect events considered as incorrect by the confidence measure) against the *false rejection rate* (the proportion of the truly correct events considered as incorrect by the confidence measure) for all possible thresholds between 0 (all events are accepted) and 1 (all events are rejected). All ROC curves are increasing functions that start at (0,0) and finish at (1,1).

In the optimal case, the point (0,1) belongs to the ROC curve (all incorrect events are detected and no correct event was rejected); thus, a ROC curve which is closer to the upper left corner of the graph represents a better confidence measure than a ROC curve which is farther from this point. The AROC measures the normalised area that covers a ROC curve, which provides a single measure of the confidence measure quality that makes results more comparable. AROC is usually normalised between 0 and 100, with 50 the baseline case. The ROC curves for FBCM and TransCM, for 2 and 3-level labels, are presented in the graphics in Figure 1. The AROC results are presented in Table 2.

From these results we can conclude that the proposed confidence measures are good enough for the annotation of dialogues. The measures behave slightly better when applied to a less complex version (with the 2-level labels), which seems reasonable since the number of different events gets reduced and the confusion gets lower. In this case, the TransCM behaves slightly better than the FBCM, but for 3-level labels differences seem not significant.

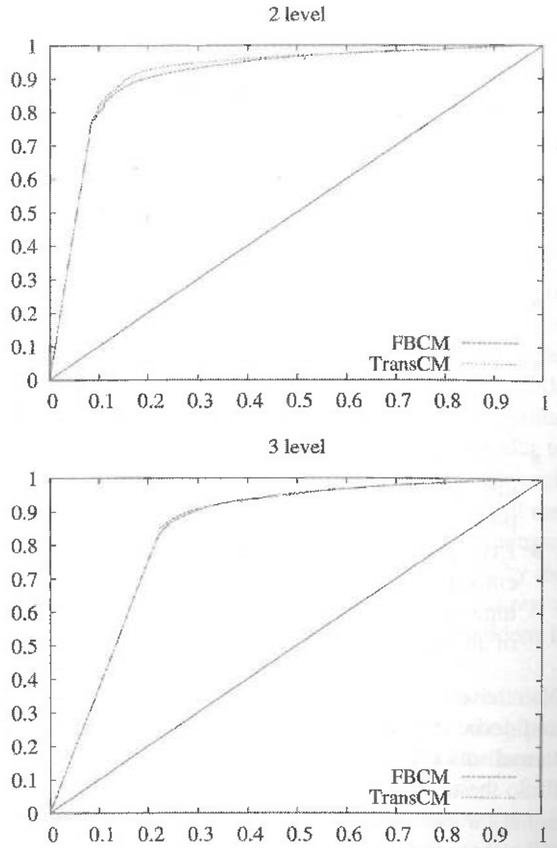

Figure 1: ROC curves for the FBCM and TransCM confidence measures, for the 2 and 3 levels label set of Dihana. Diagonal line is the baseline. The NGT model was a 3-gram and the DA output language model was a 4-gram.

Although it seems strange that TransCM behaves better than FBCM, the explanation is given by the implementation of the NGT search. Since the search space is very large, the NGT search applies intensive beam search during the process; this makes the $\alpha$ and $\beta$ computations to be inaccurate, since the beam process affects the computation of the "real" forward and backward measures (specially for $\beta$, that can be only computed when the search is finished). Thus, the inclusion of the backward probability may distort the real confidence of the local hypothesis; the forward and transition probability can be computed as the search tree is built, but when the beam condition is applied the search starts from a single branch (the current best solution) and they can locally lose precision.

## 6. Conclusions and future work

In this work we presented a proposal on confidence measures that was adapted and implemented in the NGT available software. We tested the measures on a medium-size task-oriented human-computer dialogue corpus. The results show that our proposal is good enough to provide an appropriate guidance to human correctors that must amend the draft annotation provided by the automatic technique. One of the proposals (TransCM) performs slightly better than the other (FBCM), and consequently it seems the most appropriate to be used in a real system.

Some future work must be completed to confirm the goodness of our proposal. At first term, the use of beam search in the NGT software restricts the computations of the forward and backward probabilities. Thus, the beam factor must be as high as possible to verify the real influence of these probabilities in the computation. However, the use of beam search or limited expansion of the tree is necessary to avoid the excessive spatial cost of the search process. Thus, the combination of confidence measures with a limited expansion (such as that proposed in (Tamarit et al., 2011)) is an interesting way to explore. Finally, experiments with more corpora are desirable to confirm the appropriateness of the proposal for data of different nature. The confidence measures can be applied in real annotation tasks and in the selection of the most informative dialogues to be annotated by Active Learning (Ghigi et al., 2011), in order to reduce the correction effort.

## Acknowledgments

Work supported by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018), iTrans2 (TIN2009-14511), MITTRAL (TIN2009-14633-C03-01) and RISE (TIN2008-04571) projects, by the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project and by the Generalitat Valenciana under grant Prometeo/2009/014 and GV/2010/067, and by the "VIDI-UPV" under grant UPV/2009/2851.

## 7. References

N. Alcácer, J. M. Benedí, F. Blat, R. Granell, C. D. Martínez, and F. Torres. 2005. Acquisition and Labelling of a Spontaneous Speech Dialogue Corpus. In *SPECOM*, pages 583–586, Greece.

J. L. Austin. 1962. *How to Do Things with Words*. Oxford University Press, London.

J. M. Benedí, E. Lleida, A. Varona, M. J. Castro, I. Galiano, R. Justo, I. López, and A. Miguel. 2006. Design and acquisition of a telephone spontaneous speech dialogue corpus in spanish: DIHANA. In *Fifth LREC*, pages 1636–1639, Genova, Italy.

H. Bunt. 1994. Context and dialogue control. *THINK Quarterly*, 3.

F. Casacuberta, E. Vidal, and D. Picó. 2005. Inference of finite-state transducers from regular languages. *Pattern Recognition*, 38(9):1431–1443.

M. G. Core and J. F. Allen. 1997. Coding dialogues with the DAMSL annotation scheme. In David Traum, editor, *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Menlo Park, California. AAAI.

P. A. Devijver. 1985. Baum's forward-backward algorithm revisited. *Pat. Rec. Letters*, 3(6):369 – 373.

J. P. Egan. 1975. *Signal Detection Theory and Roc Analysis*. Academic Press.

M. Fraser and G. Gilbert. 1991. Simulating speech systems. *Comp. Speech Lang.*, 5:81–99.

T. Fukada, D. Koll, A. Waibel, and K. Tanigaki. 1998. Probabilistic dialogue act extraction for concept based multilingual translation systems. In *Proceedings of ICSLP*, volume 6, pages 2771–2774.

F. Ghigi, V. Tamarit, C.-D. Martínez-Hinarejos, and J.-M. Benedí. 2011. Active learning for dialogue act labelling. In *Proceedings of IbPRIA*, pages 652–659, Las Palmas de Gran Canaria, Jun. Springer.

H. Jiang. 2005. Confidence measures for speech recognition: A survey. *Speech Comm.*, 45(4):455 – 470.

G. G. Lee, J. Mariani, W. Minker, and S. Nakamura. 2010. *Spoken Dialogue Systems for Ambient Environments*, volume 6392 of *LNCS*. Springer Verlag, Heidelberg.

C. D. Martínez-Hinarejos, J. M. Benedí, and R. Granell. 2008. Statistical framework for a spanish spoken dialogue corpus. *Speech Communication*, 50:992–1008.

C.-D. Martínez-Hinarejos, V. Tamarit, and J.-M. Benedí. 2009. Improving unsegmented dialogue turns annotation with N-gram transducers. In *Proceedings of PACLIC23*, volume 1, pages 345–354, Hong Kong, December. City University of Hong Kong Press.

R. San-Segundo, B. Pellom, K. Hacioglu, W. Ward, and J. M. Pardo. 2001. Confidence measures for spoken dialogue systems. In *ICASSP*, volume 1, pages 393–396, Los Alamitos, CA, USA. IEEE Computer Society.

R. Sánchez-Sáez, J.-A. Sánchez, and J.-M. Benedí. 2009. Statistical confidence measures for probabilistic parsing. In *Proceedings of RANLP'09*, pages 388–392, Borovets, Bulgaria, September.

S. Seneff and J. Polifroni. 2000. Dialogue management in Mercury flight reservation system. In *ANLP-NAACL*, pages 1–6.

A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer. 2000. Dialogue act modelling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):1–34.

V. Tamarit, C.-D. Martínez-Hinarejos, and J.-M. Benedí, 2011. *Spoken Dialogue Systems Technology and Design*, chapter On the Use of N-gram Transducers for Dialogue Annotation, pages 255–276. Springer.

N. Ueffing, K. Macherey, and H. Ney. 2003. Confidence measures for statistical machine translation. In *Proc. MT Summit IX*, pages 394–401. Springer-Verlag.

M. Walker and R. Passonneau. 2001. DATE: A dialogue act tagging scheme for evaluation of spoken dialogue systems. In *HLT'01: Proc. of the 1st Int. Conf. on Human language technology*, pages 1–8, San Diego.

F. Wessel, R. Schlüter, K. Macherey, and H. Ney. 2001. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9:288–298.

J. D. Williams and S. Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393 – 422.