Conditional use of Word Lattices, Confusion Networks and 1-best string hypotheses in a Sequential Interpretation Strategy

Bogdan Minescu[1], *Géraldine Damnati*[1], *Fréderic Béchet*[2], *Renato De Mori*[2]

[1] France Télécom R&D - TECH/SSTP/RVA, 2 av. Pierre Marzin, 22300 Lannion, France
[2] LIA - University of Avignon, BP1228, 84911 Avignon cedex 09 France
{bogdan.minescu,geraldine.damnati}@orange-ftgroup.com
{frederic.bechet,renato.demori}@univ-avignon.fr

Abstract

Within the context of a deployed spoken dialog service, this study presents a new interpretation strategy based on the sequential use of different ASR output representations: 1-best strings, word lattices and confusion networks. The goal is to reject as early as possible in the decoding process the non-relevant messages containing non-speech or out-of-domain content. This is done through the 1-pass of the ASR decoding process thanks to specific acoustic and language models. A confusion network (CN) is then calculated for the remaining messages and another rejection process is applied with the confidence measures obtained in the CN. The messages kept at this stage are considered relevant; therefore the search for the best interpretation is applied to a richer search space than just the 1-best word string: either the whole CN or the whole word lattice. An improved, SLU oriented, CN generation algorithm is also proposed that significantly reduces the size of the CN obtained while improving the recognition performance. This strategy is evaluated on a large corpus of real users' messages obtained from a deployed service.

Index Terms: Spoken Language Understanding, Confusion Networks, Lattice Decoding, Decision Strategy.

1. Introduction

In dialogue, interpretation goes through a sequence of phases which are controlled by imperfect knowledge and process data which may be incorrect. For this reason, it is important to reduce the probability of errors as early as possible.

Experience with the system described in [1] has shown that there are segments in a speech message which do not convey any concept in the application domain (e.g. comments made by the speakers about the service). Sophisticated concept detection methods, like those processing word lattices [7], may hypothesize incorrect domain concepts in these irrelevant segments. Such insertion errors are more frequent if the irrelevant segments are long.

This type of insertion error is very costly for a dialogue service as it might lead the system on a wrong dialog path. Detecting and rejecting these irrelevant segments is usually done thanks to confidence measures. These confidence measures can be obtained from the posterior probabilities, computed with acoustic and language models, of the words supporting the interpretation [3] they can rely on a set of features related to parsing [6] they can also contain contextual features from the dialogue as discussed in [5].

Most of these confidence features rely on word lattices. This post-processing phase is very time consuming when dealing with large word lattices. As time efficiency is one of the key parameters in a dialogue system, the use of such post-processes should be limited as much as possible to the portions of signal that are likely to contain relevant information. Using the principles described above, the interpretation strategy proposed in this paper is based on a sequential strategy using acoustic and language models to detect irrelevant segments at an early stage in the decision process. The remaining segments are converted into Confusion-Network (CN) structures that are used to reduce the search space of the word lattices and compute confidence scores for the words and concepts output. The advanced concept detection method is only applied to the portions labeled with high confidence scores.

Section 2 introduces this strategy, section 3 presents the CN algorithm specifically designed for this study and Section 4 presents the experimental setup and the results obtained on a corpus collected on a widely deployed spoken dialogue service by France Telecom R&D called the FT 3000 service.

2. Decision strategy

2.1. Characterizing audio messages

When dealing with a real service it is necessary to consider all the detections that are submitted to the speech recognition system. The content of audio messages collected through a telephone spoken dialogue service can be of various natures. This includes Non-Speech detections that may be wrongly inserted by the Noise/Speech Detection module.

In order to analyze more precisely the behavior of a system we make further distinction for speech detections. The first distinction concerns In-Domain (ID) utterances and Out-of-Domain (OOD) utterances.

OOD utterances can be composed of comments from the users (the user talks to himself or insults the system…), speech to somebody else than the service or irrelevant speech.

ID utterances are further decomposed into two categories referring to their consistency with the Spoken Language Understanding (SLU) module of the service. Actually we make a distinction between ID utterances that can be associated to an interpretation rule and the other ones that have no meaning according to the interpretation rules of the system. We can find in this last set incomplete messages, misunderstandings by the users or requests not covered by the interpretation rules.

This leads to the following categorization of corpora:

- Non-Speech detections (C1)
- Out-of-Domain speech (C2)
- In-Domain speech without interpretation (C3)
- In-Domain speech with interpretation (C4)

This work is supported by the 6th Framework Research Programme of the European Union (EU), Project LUNA, IST contract no 33549. The authors would like to thank the EU for the financial support. For more information about the LUNA project, please visit the project home-page, www.ist-luna.eu

The ASR model used in the first pass includes a rejection model trained on Non-Speech utterances. The Language Model (LM) of the first pass also includes a specific sub-LM that is dedicated to the detection of comments as presented in [1]. In order to further address the specificities of a real database, a general strategy is proposed in section 2.4 that aims at correctly rejecting utterances from C1, C2 and C3 while recognizing appropriately utterances from C4.

2.2. ASR output

When processing an audio message, the ASR module can output a word lattice as well as a 1-best string word hypothesis. This word lattice can be further processed and formatted into a Confusion-Network structure. Confusion network (CN) structure is a sequence of chunks defined by a pair of adjacent states. Each chunk, referred to as class, contains one or more competing word hypothesis denoted by corresponding links and there associated probabilities. A path is thus formed by choosing one link from each class. Given the CN structure, its main advantage is to provide to further language processing modules a reduced and structured set of hypothesis. The extraction of the best hypothesis, also called consensus hypothesis has already been shown to outperform the 1-best recognition hypothesis and the word posterior probabilities attached to each word in a confusion network are a very effective confidence measure [8].

2.3. Interpretation process

The interpretation process, as presented in [1], is based on a 2-step process: firstly detecting basic concepts thanks to a Finite-State-Machine approach and an HMM tagger; secondly composing these concepts thanks to logical rules also encoded as FSMs in order to output an interpretation of the utterance.

This 2-step process can be applied to a word string or to a set of strings encoded as FSMs (representing a word lattice or a confusion network). It has been shown in previous studies [7] that coupling the search for the best sequence of concepts and the best sequence of words through the process of word lattices can improve the interpretation performance. However using such lattices has two drawbacks:

- It is time consuming to process lattices, especially in the case of noisy messages that produce very big lattices;
- In the case of non-relevant messages corresponding to the sets C1 C2 and C3, there is a higher risk of False Alarm detection when processing lattices compared to processing a single string, as is it almost always possible to find an interpretation in a word lattice.

2.4. Strategy

From all the previous considerations, we propose the following interpretation strategy:

1. If the 1st-pass ASR model rejects the utterance, (either because of the acoustic Non-Speech detection model or because of the OOD detection LM) the utterance is considered as belonging to C1 or C2 and is directly rejected.
2. Otherwise, for each of the remaining utterances a CN is built from the corresponding WL and the consensus hypothesis is filtered according to a threshold on the confidence measures of words and then processed by the interpretation strategy; if no *reliable* interpretation is found, the message is considered as belonging to C3 and is rejected.
3. Finally the messages kept are considered as belonging to set C4; the interpretation strategy is then applied to the whole CN in order to refine the interpretation obtained on the consensus hypothesis or even to the initial word lattice if a higher precision in the detection is required

Before evaluating the strategy in section 4, the next section presents in more details the construction of CNs that has been optimized for the purpose of the study.

3. Confusion networks

In [4], a novel algorithm for the generation of confusion networks (CN) has been proposed. However, its high time complexity makes it difficult to use in real-time applications, like vocal services or call classification [8]. An alternative CN generation algorithm has been proposed in [2] in order to reduce the time complexity from which we have derived our own algorithm as explained in the next sections.

The pivot algorithm, as described in [2] consists of a clustering procedure based on an initial baseline path. This baseline path is chosen here as the shortest path in the original lattice. The pivot baseline is divided in distinct adjacent classes, so that each word forms a new class. Time delimitation is inherited from the shortest path in the lattice. The algorithm consists then of mapping each transition onto the pivot either by inserting links in one of the existing classes or by creating a new class that will be inserted in the alignment. The mapping process is guided by two parameters:

- The time overlap between the transition and the existing classes in the network
- The precedence order between the transition and the links contained in the best overlapping class.

Thus, for each transition in the topologically ordered word lattice, the algorithm searches for the best time overlapping class in the network. If the precedence order condition is satisfied, the transition is inserted in the class by either summing over the posterior probability of a link carrying the same word or by creating a new link. Otherwise, a new class is created and inserted in the network preceding the best overlapping class.

3.1. Analysis of the base-line pivot algorithm

Speech recognition systems encounter more difficulties when trying to recognize short words as compared to longer words.

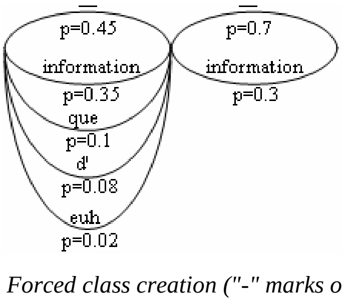

Figure 1: *Forced class creation ("-" marks omission).*

In a word lattice, the ASR system tends to generate a large number of hypotheses of the same word with different lengths, start frames and acoustic scores. It is frequent that word hypotheses having significantly different time lengths are grouped in the same class, with short words becoming possible alternatives to much longer words. This will force transitions representing good alternatives for the words of a class to create a new class due to the precedence condition.

This leads to the situation in Figure 1, with a multiplication of the same word hypothesis over more than one class and thus triggering an omission in the consensus hypothesis due to a higher omission transition posterior probability. The unusually elevated number of classes per word of reference is also a consequence of this phenomenon.

Following these observations we propose two modifications. The first one, detailed in section 3.2, is a modification of the mapping process. The second one, detailed in section 3.3, consists of a pruning stage over the final network aiming for significant size reduction.

3.2. A new concept oriented mapping technique

The initial pivot algorithm analyzes and maps transitions in a topological order. We propose a mapping technique [4] which takes into account the word carried by a transition. We define two different categories for the words in the lexicon:

- Non-empty words: associated with a concept rule.
- Empty words: not associated to any of the concept rules used by the SLU module.

The mapping process consists of three different steps. The transitions are analyzed in a topological order and inserted in the network only if they meet the criteria.

- **Step 1:** Only transitions carrying one of the words of the pivot baseline will be inserted in the corresponding classes if the precedence and time overlap conditions are satisfied. No new classes are created at this stage.
- **Step 2:** A transition carrying a non-empty word will be inserted in the network only if it conveys the same concept as one of the pivot baseline words. The precedence and time overlap conditions must be satisfied and no new classes is created.
- **Step 3:** All the remaining transitions carrying non-empty words will be inserted in the network. If the precedence condition is not satisfied for the best time overlapping class, than a new class is created and inserted in the network.

We have observed that by using only the non-empty words we obtain better performances. After Step 3, the remaining transitions carrying empty words are thus ignored by the mapping process.

3.3. Posterior pruning of final classes

In many classes in the resulting CNs, as explained in section 3.1, the omission transition carries the highest probability. Moreover, it is frequently observed that in such classes, the second best word hypothesis has a very low probability compared to the omission probability. We propose a pruning step applied to the final classes of CNs that eliminates from the network the classes where the omission has the highest posterior probability and where the ratio between its probability and the second best probability is greater than a fixed threshold. Given the CN structure, this algorithm modification has no influence on the consensus hypothesis.

3.4. Impact on CN performances

Optimizations on the CN algorithm have been performed on a development corpus consisting of 2005 utterances collected from the FT 3000 service. Table 1. shows the overall impact of the proposed modification in section 3.2. on the performance of the CN generation algorithm. The concept mapping CN algorithm yields a 3% improvement of the WER over the 1-best word lattice and the baseline pivot CN algorithm. Event though the size of the baseline pivot generated CNs is less than 2% of the size of the word lattices, the average number of classes per word of reference is unusually high. Using the concept mapping CN algorithm we manage to reduce it significantly and yield an overall size reduction of the CNs of 50%.

The degradation of the Oracle WER, when using the concept mapping CN algorithm, is a normal consequence of deliberately eliminating empty words at step 3 of the algorithm. However, this does not affect the performance of our system as we have eliminated only words which are not relevant for the interpretation process. Pruning full classes from a CN leads to a degradation of the Oracle WER as words are eliminated from the network. The threshold used for the ratio between the probability of the omission and the second best probability in the class should then be a compromise between CN size reduction and Oracle accuracy degradation. Experiments showed that the optimal value for the threshold is 10. An 80% reduction of the size of the pruned CN is achieved with les than 2% degradation of the Oracle WER.

	1-best WER	Oracle WER	Classes / Words of reference	Words / Class
Word Lattice	45.5%	19.4%	-	-
Baseline pivot CN	45.2%	9.4%	24.3	8.7
Concept mapping CN	42.5%	21.8%	14.2	6.3
Concept mapping CN + class pruning	42.5%	23.1%	1.5	8.3

Table 1. *Size and performance of CN depending on the generation algorithm*

4. Experiments and data description

4.1. Data description

The corpus we are using for this study has been obtained from the 3000 service [1], France Telecom's voice agency using natural language technologies. In order to evaluate the system on a realistic database, we have gathered for our test corpus a set of real dialogues collected from the deployed system on a two-weeks time period. The test corpus contains 6501 utterances (3200 dialogs). Considering the categorization of section 2.1, it can be decomposed as described in Table 2.

Category	# utterances
C1: Non-Speech detections	1333
C2: Out-of-Domain speech	674
C3: In-Domain speech without interpretation	355
C4: In-Domain speech with interpretation	4139
Total	6501

Table 2: *Description of the test corpus*

4.2. Experimental framework

In order to perform an evaluation that reflects as much as possible the behavior of the system from the user's point of view, we focus on the evaluation of the Interpretation Error Rate (IER). Interpretations are the input of the Dialogue Manager. The reference manual transcription of an utterance is submitted to the SLU module, leading to a reference interpretation which is compared to the interpretation hypothesized by our strategy. An interpretation generated by the SLU module implemented in the 3000 service refers to a composition of attribute-value pairs. An interpretation is considered correct if all the attribute-value pairs and their relations are correct. Thus, three types of errors may occur:

- *False Alarms* (*FA*) when an interpretation is output by the system while the reference interpretation is empty.
- *Substitutions* (*Sub*) when a hypothesized interpretation is different from the reference interpretation.
- *False Rejections* (*FR*) when no interpretation is hypothesized while the reference interpretation is not empty.

FA can occur either on Non-Speech detections (C1), Out-of-Domain utterances (C2) or In-Domain utterances without interpretations (C3). *Sub* and *FR* are two types of errors that can occur on In-Domain utterances with interpretations (C4).

The overall IER is obtained by summing up the different errors (*FA*+*Sub*+*FR*) and dividing by the total amount of non-empty reference interpretation.

4.3. Experiments

Before evaluating the general strategy proposed in section 2.4, a first set of experiments aims at showing the performance of each of the possible decoding schemes when applied separately on the whole corpus. Five outputs are compared in Table 3 in terms of errors at the interpretation level. *WL 1-best* is the interpretation obtained on the best path of the Word Lattice. *CN_baseline 1-best* is the interpretation obtained on the Consensus Hypothesis of the baseline pivot CN algorithm and *CN 1-best* is the interpretation obtained on the Consensus Hypothesis of the proposed algorithm after applying a confidence measure threshold on the hypothesized words. *WL Decoding* and *CN Decoding* represent the interpretations obtained with an integrated search (word+interpretation) on the whole word lattices and confusion networks, as presented in [7].

	WL 1-best	CN_baseline 1-best	CN 1-best	WL Decoding	CN Decoding
FA on C1	6.5 %	3.8%	**2.6 %**	22.8 %	20.1 %
FA on C2	7.8 %	6.2%	**5.3 %**	13.0 %	13.7 %
FA on C3	2.9 %	2.5%	**2.3 %**	6.3 %	7.0 %
Sub+FR on C4	8.8 %	11.5%	10.6 %	**6.5 %**	8.6 %

Table 3: *Detailed errors on the whole corpus when applying the different methods independently.*

As expected, using richer search space like word lattices improves the interpretation accuracy on the set C4 (from 8.8% to 6.5% in interpretation errors) but increases at the same time the False Alarm measure (up to 22.8% on C1). On the contrary the Consensus Hypothesis gives the best results for reducing the amount of FA errors on the sets C1, C2 and C3.

The strategy proposed in section 2.4 is evaluated in Table 4 through two implementations: *Strat1* and *Strat2*. The rejection process, based on ASR models and CNs as presented in 2.4., is identical for both. They differ only on the search space used for finding the best interpretations in the messages classified as relevant: whole CN for *Strat1* and whole WL for *Strat2*.

These strategies are compared to the baseline one that uses only the 1-best string of the WL in order to either reject a message or look for the best interpretation. We can see that a very significant improvement is obtained by using CN instead of the 1-best hypothesis. This improvement is mainly due to the rejection strategy based on the confidence measures of the CNs, which yields 8.4% absolute reduction in the overall *FA* rate while degrading the *FR* rate by only 2.5%. Using the CN search space in *Strat 1* reduces the *Sub* rate. By using a richer search space for the messages considered as relevant, another small improvement is achieved with WL on the *Sub* rate.

Moreover, only 75% of the CNs or WL are explored when using either of the strategies, resulting in a reduced decoding time.

total	Baseline (1-best)	*Strat1* (CN)	*Strat2* (WL)
FA	17.2 %	8.8 %	8.8 %
Sub	6.1 %	5.6 %	4.1 %
FR	2.7 %	5.2 %	5.2 %
IER	**26.0 %**	**19.6 %**	**18.1 %**

Table 4: *Total Interpretation Error Rate (IER) with the baseline system and the 2 strategies proposed.*

5. Conclusion

This study presents results for a new sequential interpretation strategy for a dialogue service. Its aim is to reject non-relevant messages using 1-best ASR and consensus hypothesis and to explore CNs or word lattices on relevant messages only. While exploring only 75% of the CNs or word lattices, the strategy yields significant IER improvement.

An improved, SLU oriented, CN generation algorithm is also proposed and results show improved WER performance and significant size reduction. The resulting CN proved to be very efficient in the general strategy.

6. References

[1] G. Damnati, F. Bechet, R. de Mori, "Spoken language understanding strategies on the France Telecom 3000 voice agency corpus", Proc. ICASSP, IV:9-12, 2007.

[2] D. Hakkani-Tur, F. Bechet, G. Riccardi and G. Tur, "Beyond ASR 1-Best: Using Word Confusion Networks for Spoken Language Understanding", Computer Speech and Language, 20(4):495-514, 2006.

[3] R. Lieb, T. Fabian, G. Ruske, M. Thomae, "Estimation of Semantic Confidences on Lattice Hierarchies", Proc. ICSLP, 569-572, 2004.

[4] L. Mangu, E. Brill, and A. Stolcke, "Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks.", Computer Speech and Language, 14(4):373-400, 2000.

[5] M. Purver, F. Ratiu, L. Cavedon, "Robust Interpretation in Dialogue by Combining Confidence Scores with Contextual Features", Proc. ICSLP, 1-4, 2006.

[6] R. Sarikaya, Y. Gao, M. Picheny and H. Erdogan, "Semantic confidence measurement for spoken dialog systems", IEEE Trans. on SAP, 13(4): 534-545, 2005.

[7] C. Servan, C. Raymond, F. Bechet and P. Nocera, "Conceptual Decoding from Word Lattices: Application to the Spoken Dialogue Corpus MEDIA", Proc. ICSLP, 1614-1617, 2006.

[8] G. Tur, D. Hakkani-Tur, G. Riccardi, " Extending boosting for call classification using word confusion networks", Proc. ICASSP, 437-440, 2004