Using Paraphrases and Lexical Semantics to Improve the Accuracy and the Robustness of Supervised Models in Situated Dialogue Systems

Claire Gardent
CNRS/LORIA, Nancy
claire.gardent@loria.fr

Lina M. Rojas Barahona
Université de Lorraine/LORIA, Nancy
lina.rojas@loria.fr

Abstract

This paper explores to what extent lemmatisation, lexical resources, distributional semantics and paraphrases can increase the accuracy of supervised models for dialogue management. The results suggest that each of these factors can help improve performance but that the impact will vary depending on their combination and on the evaluation mode.

1 Introduction

One strand of work in dialog research targets the rapid prototyping of virtual humans capable of conducting a conversation with humans in the context of a virtual world. In particular, question answering (QA) characters can respond to a restricted set of topics after training on a set of dialogs whose utterances are annotated with dialogue acts (Leuski and Traum, 2008).

As argued in (Sagae et al., 2009), the size of the training corpus is a major factor in allowing QA characters that are both robust and accurate. In addition, the training corpus should arguably be of good quality in that (i) it should contain the various ways of expressing the same content (paraphrases) and (ii) the data should not be skewed. In sum, the ideal training data should be large (more data is better data) ; balanced (similar amount of data for each class targeted by the classifier) and varied (it should encompass the largest possible number of paraphrases and synonyms for the utterances of each class).

In this paper, we explore different ways of improving and complementing the training data of a supervised QA character. We expand the size and the quality (less skewed data) of the training corpus using paraphrase generation techniques. We compare the performance obtained on lemmatised vs. non lemmatised data. And we investigate how various resources (synonym dictionaries, WordNet, distributional neighbours) can be used to handle unseen words at run time.

2 Related work

Previous work on improving robustness of supervised dialog systems includes detecting and handling out of domain utterances for generating feedback (Lane et al., 2004) ; using domain-restricted lexical semantics (Hardy et al., 2004) ; and work on manual data expansion (DeVault et al., 2011). Our work follows up on this research but provides a systematic investigation of how data expansion, lemmatisation and synonym handling impacts the performance of a supervised QA engine.

3 Experimental Setup

We run our experiments on a dialog engine developed for a serious game called *Mission Plastechnologie*. In this game, the player must interact with different virtual humans through a sequence of 12 subdialogs, each of them occurring in a different part of the virtual world.

Training Data. The training corpus consists of around 1250 Human-Human dialogues which were manually annotated with dialog moves. As the following dialog excerpt illustrates, the dialogs are conducted in French and each dialog turn is manually annotated using a set of 28 dialog acts. For

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 808–813,
Seattle, Washington, USA, 18-21 October 2013.

a more detailed presentation of the training corpus and of the annotation scheme, the reader is referred to (Rojas-Barahona et al., 2012a)

dialog : 01_dialogDirecteur-Tue Jun 14 11 :04 :23 2011

>M.Jasper : Bonjour, je suis M.Jasper le directeur. || ***greet***

(Hello, I am the director, Mr. Jasper.)

>M.Jasper : Qu'est-ce que je peux faire pour vous ? || ***ask(task(X))***

(What can I do for you ?)

>Lucas : je dois sauver mon oncle || ***first_step***

(I must rescue my uncle)

>M.Jasper : Pour faire votre manette, il vous faut des plans. Allez voir dans le bureau d'études, ils devraient y être. || ***inform(do(first_step))***

(To build the joystick you will need the plans. You will find them in the Designing Office.)

>M.Jasper : Bonne Chance ! || ***quit***

(Good Luck !)

Dialog Systems For our experiments, we use a hybrid dialog system similar to that described in (Rojas Barahona et al., 2012b; Rojas Barahona and Gardent, 2012). This system combines a classifier for interpreting the players utterances with an information state dialog manager which selects an appropriate system response based on the dialog move assigned by the classifier to the user turn. The classifier is a logistic regression classifier [1] which was trained for each subdialog in the game. The features used for training are the set of content words which are associated with a given dialog move and which remain after TF*IDF [2] filtering. Note that in this experiment, we do not use contextual features such as the dialog acts labeling the previous turns. There are two reasons for this. First, we want to focus on the impact of synonym handling, paraphrasing and lemmatisation on dialog management. Removing contextual features allows us to focus on how content features (content words) can be improved by these mechanisms. Second, when evaluating on the H-C corpus (see below), contextual features are often incorrect (because the system might incorrectly interpret and thus label a user turn). Excluding contextual features from training allows for a fair comparison between the H-H and the H-C evaluation.

Test Data and Evaluation Metrics We use accuracy (the number of correct classifications divided by the number of instances in the testset) to measure performance and we carry out two types of evaluation. On the one hand, we use 10-fold cross-validation on the EmoSpeech corpus (H-H data). On the other hand, we report accuracy on a corpus of 550 Human-Computer (H-C) dialogues obtained by having 22 subjects play the game against the QA character trained on the H-H corpus. As we shall see below, performance decreases in this second evaluation suggesting that subjects produce different turns when playing with a computer than with a human thereby inducing a weak out-of-domain effect and negatively impacting classification. Evaluation on the H-H corpus therefore gives a measure of how well the techniques explored help improving the dialog engine when used in a real life setting.

Correspondingly, we use two different tests for measuring statistical significance. In the H-H evaluation, significance is computed using the Wilcoxon signed rank test because data are dependent and are not assumed to be normally distributed. When building the testset we took care of not including paraphrases of utterances in the training partition (for each paraphrase generated automatically we keep track of the original utterance), however utterances in both datasets might be generated by the same subject, since a subject completed 12 distinct dialogues during the game. Conversely, in the H-C evaluation, training (H-H data) and test (H-C data) sets were collected under different conditions with different subjects therefore significance was computed using the McNemar sign-test (Dietterich, 1998).

4 Paraphrases, Synonyms and Lemmatisation

We explore three main ways of modifying the content features used for classification : lemmatising the training and the test data ; augmenting the training data with automatically acquired paraphrases ; and substituting unknown words with synonyms at run time.

Lemmatisation We use the French version of Treetagger [3] to lemmatise both the training and the test data. Lemmas without any filtering were used

1. We used MALLET (McCallum, 2002) for the LR classifier with L1 Regularisation.

2. TF*IDF = Term Frequency*Inverse Document Frequency

3. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

to train classifiers. We then compare performance with and without lemmatisation. As we shall see, the lemma and the POS tag provided by TreeTagger are also used to lookup synonym dictionaries and EuroWordNet when using synonym handling at run time.

Paraphrases : (DeVault et al., 2011) showed that enriching the training corpus with manually added paraphrases increases accuracy. Here we exploit automatically acquired paraphrases and use these not only to increase the size of the training corpus but also to better balance it [4]. We proceed as follows.

First, we generated paraphrases using a pivot machine translation approach where each user utterance in the training corpus (around 3610 utterances) was translated into some target language and back into French. Using six different languages (English, Spanish, Italian, German, Chinese and Arabian), we generated around 38000 paraphrases. We used Google Translate API for translating.

Category	Train Instances	Balanced Instances
greet	24	86
help	20	82
yes	92	123
no	55	117
ack	73	135
other	27	89
quit	38	100
find_plans	115	146
job	26	88
staff	15	77
studies	20	82
security_policies	24	86
μ	44.08	100.92
σ	$\pm$32.68	$\pm$23.32

TABLE 1: Skewed and Balanced Data on a sample subdialog. The category with lowest number of paraphrases is `greet`, with 62 paraphrases, hence $l_p = 62$. All categories were increased by 62 except `find_plans` and `yes` that were increased by half : 31.

Second, we eliminate from these paraphrases, words that are likely to be incorrect lexical translations by removing words with low normalized term frequency (< 0.001) across translations i.e., lexical translations given by few translations and/or translation systems. We then preprocessed the paraphrases in the same way the utterances of the initial training corpus were preprocessed i.e., utterances were unaccented, converted to lower-case and stop words were removed, the remaining words were filtered with TF*IDF. After preprocessing, duplicates were removed.

4. The Emospeech data is highly skewed with some classes being populated with many utterances and others with few.

```
Algorithm extendingDataWithParaphrases(trainingset ts)
1. Let c be the set of categories in ts.
2. μ be the mean of train instances per category
3. σ be the standard deviation of train instances per category
4. Let Np_c be the number of paraphrases per category
5. Let l_p ← min Np_cj
6. Repeat
7.   set i ← 0
8.   Ninst_ci be the number of instances per category c_i
9.   d_i ← Ninst_ci − μ
10.  if d_i < σ then
11.    Ninst_ci ← l_p
12.  else
13.    Ninst_ci ← l_p/2
14.  end if
15.  set i←i+1
16.  if i>‖c‖ then
17.   terminate
18. end
```

FIGURE 1: Algorithm for augmenting the training data with paraphrases.

Third, we added the paraphrases to the training data seeking to improve the balance between dialog moves per dialog, as shown in Figure 1. To this end, we look for the category c with the lowest number of paraphrases l_p (line 5). We then compute the deviation d_i for each dialog move c_i from the mean μ in the original training set (line 9). If the deviation d_i is lower than the standard deviation then we add l_p number of paraphrases instances (line 11). Conversely, if d_i is higher than the standard deviation, we reduce the number of instances to be added by half $\frac{l_p}{2}$ (line 13). Table 1 shows the original and the extended training data for the third sub-dialog in the Emospeech game. In this dialogue the player is supposed to ask information about the joystick plans ($find\_plans$, which is the mandatory goal). The categories cover mandatory and optional goals and general dialogue acts, such as greetings, asking for help, confirm and disconfirm, acknowledgment and out of topic questions (i.e. other).

Substituting Synonyms for Unknown Words A word is unknown, if it is a well-formed French

word [5] and if it does not appear in the training corpus. Conversely, a word is known if it is not unknown.

When an unknown word w is detected in a player utterance at runtime, we search for a word w' which occurs in the training data and is either a synonym of w or a distributional neighbour. After disambiguation, we substitute the unknown word for the synonym.

To identify synonyms, we make use of two lexical resources namely, the French version of EuroWordNet (EWN) (Vossen, 1998), which includes 92833 synonyms, hyperonyms and hyponyms pairs, and a synonym lexicon for French (DIC) [6] which contains 38505 lemmas and 254149 synonym pairs. While words are categorised into Noun, Verbs and Adjectives in EWN, DIC contains no POS tag information.

To identify distributional neighbours, we constructed semantic word spaces for each subdialog in the EmoSpeech corpus [7] using random indexing (RI) [8] on the training corpus expanded with paraphrases. Using the cosine measure as similarity metrics, we then retrieve for any unknown word w, the word w' which is most similar to w and which appear in the training corpus.

For lexical disambiguation, two methods are compared. We use the POS tag provided by TreeTagger. In this case, disambiguation is syntactic only. Or we pick the synonym with highest probability based on a trigram language model trained on the H-H corpus [9].

5 Results and Discussion

Table 2 summarises the results obtained in four main configurations : (i) with and without paraphrases ; (ii) with and without synonym handling ; (iii) with and without lemmatisation ; and (iv) when combining lemmatisation with synonym handling. We also compare the results obtained when evaluating using 10-fold cross validation on the training data (H-H dialogs) vs. evaluating the performance of the system on H-C interactions.

Overall Impact The largest performance gain is obtained by a combination of the three techniques explored in this paper namely, data expansion, synonym handling and lemmatisation (+8.9 points for the cross-validation experiment and +2.3 for the H-C evaluation).

Impact of Lexical Substitution at Run Time Because of space restrictions, we do not report here the results obtained using lexical resources without lemmatisation. However, we found that lexical resources are only useful when combined with lemmatisation. This is unsurprising since synonym dictionaries and EuroWordNet only contain lemmas. Indeed when distributional neighbours are used, lemmatisation has little impact (e.g., 65.11% using distributional neighbours without lemmatisation on the H-H corpus without paraphrases vs. 66.41% when using lemmatisation).

Another important issue when searching for a word synonym concerns lexical disambiguation : the synonym used to replace an unknown word should capture the meaning of that word in its given context. We tried using a language model trained on the training corpus to choose between synonym candidates (i.e., selecting the synonym yielding the highest sentence probability when substituting that synonym for the unknown word) but did not obtain a significant improvement. In contrast, it is noticeable that synonym handling has a higher impact when using EuroWordNet as a lexical resource. Since EuroWordNet contain categorial information while the synonym dictionaries we used do not, this suggests that the categorial disambiguation provided by TreeTagger helps identifying an appropriate synonym in EuroWordNet.

Finally, it is clear that the lexical resources used for this experiment are limited in coverage and quality. We observed in particular that some words which are very frequent in the training data (and thus which could be used to replace unknown words) do not occur in the synonym dictionaries. For instance when using paraphrases and dictionaries (fourth row and

5. A word is determined to be a well-formed French word if it occurs in the LEFFF dictionary, a large-scale morphological and syntactic lexicon for French (Sagot, 2010)

6. DICOSYN (http ://elsap1.unicaen.fr/dicosyn.html).

7. We also used distributional semantics from the Gigaword corpus but the results were poor probably because of the very different text genre and domains between the the Gigaword and the MP game.

8. Topics are Dialog acts while documents are utterances ; we used the S-Space Package `http://code.google.com/p/airhead-research/wiki/RandomIndexing`

9. We used SRILM (`http://www.speech.sri.com/projects/srilm`)

H		Lemmatisation			
H-H	Orig.	Lemmas	+EWN	+DIC	+RI
Orig.	$65.70\% \pm 5.62$	$\mathbf{66.04\% \pm 6.49}$	$68.17\% \pm 6.98$	$67.92\% \pm 4.51$	$66.83\% \pm 5.92$
Parap.	$70.89\% \pm 6.45$	$\mathbf{74.31\% \pm 4.78}$*	$\mathbf{74.60\% \pm 5.99}$*	$\mathbf{73.07\% \pm 7.71}$*	$\mathbf{72.63\% \pm 5.82}$*
H-C	Orig.	Lemmas	+EWN	+DIC	+RI
Orig.	$59.71\% \pm 16.42$	$59.88\% \pm 7.19$	$61.14\% \pm 16.65$	$61.41\% \pm 16.59$	$60.75\% \pm 17.39$
Parap.	$59.82\% \pm 15.53$	$59.48\% \pm 14.02$	$\mathbf{61.70\% \pm 14.09}$*	$\mathbf{62.01\% \pm 14.37}$*	$61.16\% \pm 14.41$*

TABLE 2: Accuracy on the H-H and on the H-C corpus. The star denotes statistical significance with the Wilcoxon test ($p < 0.005$) used for the HH corpus and the McNemar test ($p < 0.005$) for the HC corpus.

fourth column in Table 2) 50% of the unknown words were solved, 17% were illformed and 33% remained unsolved. To compensate this deficiency, we tried combining the three lexical resources in various ways (taking the union or combining them in a pipeline using the first resource that would yield a synonym). However the results did not improve and even in some cases worsened due probably to the insufficient lexical disambiguation. Interestingly, the results show that paraphrases always improves synonym handling presumably because it increases the size of the known vocabulary thereby increasing the possibility of finding a known synonym.

In sum, synonym handling helps most when (i) words are lemmatised and (ii) unknown words can be at least partially (i.e., using POS tag information) disambiguated. Moreover since data expansion increases the set of known words available as potential synonyms for unknown words, combining synonym handling with data expansion further improves accuracy.

Impact of Lemmatisation When evaluating using cross validation on the training corpus, lemmatisation increases accuracy by up to 3.42 points indicating that unseen word forms negatively impact accuracy. Noticeably however, lemmatisation has no significant impact when evaluating on the H-C corpus. This in turn suggests that the lower accuracy obtained on the H-C corpus results not from unseen word forms but from unseen lemmas.

Impact of Paraphrases On the H-H corpus, data expansion has no significant impact when used alone. However it yields an increase of up to 8.27 points and in fact, has a statistically significant impact, for all configurations involving lemmatisation. Thus, data expansion is best used in combination with lemmatisation and their combination permits creating better, more balanced and more general training data. On the H-C corpus however, the impact is negative or insignificant suggesting that the decrease in performance on the H-C corpus is due to content words that are new with respect to the training data i.e., content words for which neither a synonym nor a lemma can be found in the expanded training data.

Conclusion

While classifiers are routinely trained on dialog data to model the dialog management process, the impact of such basic factors as lemmatisation, automatic data expansion and synonym handling has remained largely unexplored. The empirical evaluation described here suggests that each of these factors can help improve performance but that the impact will vary depending on their combination and on the evaluation mode. Combining all three techniques yields the best results. We conjecture that there are two main reasons for this. First, synonym handling is best used in combination with POS tagging and lemmatisation because these supports partial lexical semantic disambiguation. Second, data expansion permits expanding the set of known words thereby increasing the possibility of finding a known synonym to replace an unknown word with.

Acknowledgments

This work was partially supported by the EU funded Eurostar EmoSpeech project. We thank Google for giving us access to the University Research Program of Google Translate.

References

David DeVault, Anton Leuski, and Kenji Sagae. 2011. Toward learning and evaluation of dialogue policies with text examples. In *12th SIGdial Workshop on Discourse and Dialogue*, Portland, OR, June.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10 :1895–1923.

Hilda Hardy, Tomek Strzalkowski, Min Wu, Cristian Ursu, Nick Webb, Alan W. Biermann, R. Bryce Inouye, and Ashley McKenzie. 2004. Data-driven strategies for an automated dialogue system. In *ACL*, pages 71–78.

Ian Richard Lane, Tatsuya Kawahara, and Shinichi Ueno. 2004. Example-based training of dialogue planning incorporating user and situation models. In *INTERSPEECH*.

Anton Leuski and David Traum. 2008. A statistical approach for text processing in virtual humans. In *Proceedings of the 26th Army Science Conference*.

Andrew Kachites McCallum. 2002. Mallet : A machine learning for language toolkit. http ://mallet.cs.umass.edu.

Lina Maria Rojas Barahona and Claire Gardent. 2012. What should I do now ? Supporting conversations in a serious game. In *SeineDial 2012 - 16th Workshop on the Semantics and Pragmatics of Dialogue*, Paris, France. Jonathan Ginzburg (chair), Anne Abeillé, Margot Colinet, Gregoire Winterstein.

Lina M. Rojas-Barahona, Alejandra Lorenzo, and Claire Gardent. 2012a. Building and exploiting a corpus of dialog interactions between french speaking virtual and human agents. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*.

Lina M. Rojas Barahona, Alejandra Lorenzo, and Claire Gardent. 2012b. An end-to-end evaluation of two situated dialog systems. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–19, Seoul, South Korea, July. Association for Computational Linguistics.

K. Sagae, G. Christian, D. DeVault, , and D.R. Traum. 2009. Towards natural language understanding of partial speech recognition results in dialogue systems. In *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Companion Volume : Short Papers*, pages 53–56.

Benoît Sagot. 2010. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

Piek Vossen, editor. 1998. *EuroWordNet : a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.