Classification and Resolution of Non-Sentential Utterances in Dialogue

Paolo Dragone*
Università di Trento

Pierre Lison**
University of Oslo

This article addresses the problems of classification and resolution of non-sentential utterances (NSUs) in dialogue. NSUs are utterances that do not have a complete sentential form but convey a full clausal meaning given the conversational context, such as "To the contrary!" or "How much?". The presented approach builds upon the work of Fernández, Ginzburg, and Lappin (2007), who provide a taxonomy of NSUs divided in 15 classes along with a small annotated corpus extracted from dialogue transcripts. The main part of this article focuses on the automatic classification of NSUs according to these classes. We show that a combination of novel linguistic features and active learning techniques yields a significant improvement in the classification accuracy over the state-of-the-art, and is able to mitigate the scarcity of labelled data. Based on this classifier, the article also presents a novel approach for the semantic resolution of NSUs in context using probabilistic rules.

1. Introduction

In dialogue, utterances do not always take the form of complete, well-formed sentences with a subject, a verb and complements. Many utterances – often called *non-sentential utterances*, or NSUs for short – are fragmentary and lack an overt predicate such as a finite verb. Consider the three following examples from the British National Corpus (with the exact corpus references indicated in brackets):

A: I thought he said fifty as well.
B: **Oh no, fifteen.** [BNC: J9A 376-377]

A: Got to do a marathon tomorrow.
B: **Why?** [BNC: G4W 274-275]

A: The Holy Spirit is one who <pause> gives us?
B: **Strength.** [BNC: HDD 277-278]

Despite their ubiquity in spontaneous conversations, the semantic content of NSUs is often difficult to extract automatically. Non-sentential utterances are indeed intrinsically dependent on the surrounding dialogue context for their interpretation – for instance, the exact meaning of the utterance "Why?" in the example above is impossible to decipher without knowing the statement that precedes it. Similarly, the utterance "Strength." only makes sense in relation to the preceding, unfinished sentence.

* Dept. of Information Engineering and Computer Science - Via Sommarive 9, 38123 Povo (TN), Italy. E-mail: paolo.dragone@unitn.it

** Institutt for informatikk, Postboks 1080 Blindern, 0316 Oslo. E-mail: plison@ifi.uio.no

The first step in the interpretation of non-sentential utterances is to detect the correct *type* of each NSU. The seminal work of Fernández, Ginzburg, and Lappin (2007) introduced a taxonomy of non-sentential utterances factored into 15 classes. They also provided a small annotated corpus based on this taxonomy, and a simple classification approach to determine the type of NSUs from contextual and linguistic features. The present article builds upon and extend their work. In particular, we demonstrate that the introduction of new linguistic features combined with the use of active learning provide a modest but significant improvement in the classification accuracy compared to their approach.

In complement to the classification of NSUs, the article also presents a proof-of-concept implementation for the semantic resolution of NSUs. The objective of semantic resolution is to infer an appropriate semantic representation of non-sentential utterances (usually expressed as a logical form) on the basis of the dialogue context. Previous work formalised this task in terms of logical rules operating on a representation of the current dialogue state (Fernández 2006; Ginzburg 2012; Schlangen 2003). Logic-based formalisms are, however, not well-suited to capture – and reason over – uncertain knowledge and non-deterministic conversational behaviours. This paper presents an alternative approach based on probabilistic rules (Lison 2015). Probabilistic rules are defined as structured mappings between logical conditions and probabilistic effects. Thanks to their use of logical abstractions, these rules are able to capture complex relational structures, but do so within a probabilistic setting that can account for uncertain, ambiguous or unforeseen inputs.

The next section reviews the key concepts behind the representation and interpretation of non-sentential utterances, and presents the corpus and taxonomy used in this work. Section 3 describes our classification approach, including the feature set and active learning strategy. Section 4 then details the empirical results on the classification task and compares them with the baseline. Section 5 presents a brief summary of the approach employed to resolve the semantic content of NSUs using probabilistic rules. Section 6 discusses the advantages and shortcomings of the presented approach in relation to previous work. Finally, Section 7 concludes this paper.

2. Background

2.1 Generalities

Non-sentential utterances (also called elliptical utterances or fragments) are utterances that "do not have the form of a full sentence according to most traditional grammars, but that nevertheless convey a complete sentential meaning, usually a proposition or a question." (Fernández 2006). The meaning of these utterances must therefore be inferred from the surrounding context.

The most important contextual factor is the *antecedent* of the NSU, which is the utterance in the dialogue history that contains its underspecified content. The antecedent is often, but not always, the immediately preceding utterance in the history. This can be illustrated with the following example:

> ANNE: Peter, would you like to talk about language register in tabloids and broadsheets?
> ANNE: I'm just gonna see if I can find a clarity index.
> PETER: **Okay.** [BNC: J40 54-56]

We can observe from short dialogue above that the answer "Okay" is a response to the request in the first utterance. In other words, the correct antecedent of "Okay" is the first utterance, not the second.

There is often a tight linguistic coupling between the NSU and its antecedent, leading to various forms of parallelism between the two utterances, as described in detail by Dalrymple, Shieber, and Pereira (1991), Fernández (2006), Ginzburg (2012). This parallelism is often phonological, where a portion of the antecedent is repeated in the non-sentential utterance:

A: Find I'm down to using it twice a week now.
B: **Twice a week.** [BNC: G5X 137-138]

Parallelism can also be of a syntactic or semantic nature, such as in the following excerpt, where the prepositional phrase "from X" from the antecedent is reused in the non-sentential utterance:

A: Can you tell me where you actually got that information from?
B: **From our wages and salary department.** [BNC: K6Y 94-95]

The parallelism between non-sentential utterances and their corresponding antecedents is especially prominent in particular NSU classes such as clarification ellipsis. As described in Section 3, this linguistic fact can be fruitfully exploited in the classification process.

2.2 Taxonomy

Fernández, Ginzburg, and Lappin (2007) provide a taxonomy of non-sentential utterances based on 15 classes, reflecting both the form and pragmatic function fulfilled by the utterance:

- *Plain acknowledgement*: Used to signal understanding or acceptance of the preceding utterance, often with words such as "yeah", "right", "mhm".

  A: I shall be getting a copy of this tape.
  B: **Right.** [BNC: J42 71-72]

- *Repeated acknowledgement*: Acknowledgement that make use of repetition or reformulation of some constituent of the antecedent.

  A: Oh so if you press enter it'll come down one line.
  B: **Enter.** [BNC: G4K 102-103]

- *Clarification ellipsis*: Used to request a clarification of some aspect of the antecedent that was not fully understood.

  A: I would try F ten.
  B: **Just press F ten?** [BNC: G4K 72-73]

- *Check question*: Used to request an explicit feedback of understanding or acceptance, usually uttered by the same speaker as the antecedent.

A: So (pause) I'm allowed to record you.
A: **Okay?** [BNC: KSR 5-6]

- *Sluice*: Used to request additional information related to the antecedent.

  A: They wouldn't do it, no.
  B: **Why?** [BNC: H5H 202-203]

- *Filler*: Used to complete a previous unfinished utterance.

  A: [...] would include satellites like erm
  B: **Northallerton.** [BNC: H5D 78-79]

- *Short Answer*: Answers to wh-questions.

  A: What's plus three times plus three?
  B: **Nine.** [BNC: J91 172-173]

- *Affirmative Answer* and *Rejection*: Used to answer polar questions using yes-words and no-words.

  A: Have you settled in?
  B: **Yes, thank you.** [BNC: JSN 36-37]

  A: (pause) Right, are we ready?
  B: **No, not yet.** [BNC: JK8 137-138]

- *Repeated Affirmative Answer*: Used to give an affirmative answer by repeating or reformulating part of the query.

  A: You were the first blind person to be employed in the County Council?
  B: **In the County Council, yes.** [BNC: HDM 19-20]

- *Helpful Rejection*: Used to correct a piece of information in the antecedent.

  A: Right disk number four?
  B: **Three.** [BNC: H61 10-11]

- *Propositional Modifier* and *Factual Modifier*: Used to add modal or attitudinal information, often with modal adverbs or factual adjectives.

  A: Oh you could hear it?
  B: **Occasionally yeah.** [BNC: J8D 14-15]

  A: You'd be there six o'clock gone mate.
  B: **Wonderful.** [BNC: J40 164-165]

- *Bare Modifier Phrase*: Modifiers that behave like non-sentential adjunct modifying a contextual utterance.

A: [...] then across from there to there.
B: **From side to side.** [BNC: HDH 377-378]

- *Conjunct*: Modifier extending a previous utterance through a conjunction.

A: I'll write a letter to Chris
B: **And other people.** [BNC: G4K 19-20]

2.3 Annotated corpus

In complement to the above taxonomy, (Fernández, Ginzburg, and Lappin 2007) also presents a small corpus of annotated NSUs extracted from dialogue transcripts of the British National Corpus (Burnard 2000). The dialogue transcripts used in the corpus study contains both two-party and multi-party conversations, and cover a wide variety of dialogue domains, from formal situations (such as interviews or seminars) to everyday conversations.

A total of about 14000 utterances from 54 files were examined by the annotators, resulting in a corpus of 1299 non-sentential utterances (9 % of the total number of utterances). Of the extracted NSUs, 1283 were successfully categorized according to the above taxonomy (98.9 % coverage). Each instance of NSU is annotated with its corresponding class and its antecedent (which is often but not always the preceding utterance). Table 1 provides the frequency of each of the 15 class in the taxonomy in the corpus.

Table 1
Taxonomy of non-sentential utterances along with constructed examples and frequencies in the annotated excerpt from the BNC corpus (Fernández, Ginzburg, and Lappin 2007). The examples are borrowed from Ginzburg (2012).

NSU Class	Example		Nb. inst.	Frequency
Plain Acknowledgement	*A: ...*	*B: mmh*	599	46.7 %
Short Answer	*A: Who left?*	*B: Bo*	188	14.7 %
Affirmative Answer	*A: Did Bo leave?*	*B: Yes*	105	8.2 %
Repeated Acknowledgment	*A: Did Bo leave?*	*B: Bo, hmm.*	86	6.7 %
Clarification Ellipsis	*A: Did Bo leave?*	*B: Bo?*	82	6.4 %
Rejection	*A: Did Bo leave?*	*B: No.*	49	3.8 %
Factual Modifier	*A: Bo left.*	*B: Great!*	27	2.1 %
Repeated Affirmative Answer	*A: Did Bo leave?*	*B: Bo, yes.*	26	2.0 %
Helpful Rejection	*A: Did Bo leave?*	*B: No, Max.*	24	1.9 %
Check Question	*A: Bo isn't here. okay?*		22	1.7 %
Sluice	*A: Someone left.*	*B: Who?*	21	1.6 %
Filler	*A: Did Bo ...*	*B: leave?*	18	1.4 %
Bare Modifier Phrase	*A: Max left.*	*B: Yesterday.*	15	1.2 %
Propositional Modifier	*A: Did Bo leave?*	*B: Maybe.*	11	0.9 %
Conjunct	*A: Bo left.*	*B: And Max.*	10	0.8 %
Total			1283	100 %

3. Approach

In addition to their corpus and taxonomy of NSUs, (Fernández, Ginzburg, and Lappin 2007) also described a simple machine learning approach to determine the NSU class from simple features. Their approach will constitute the baseline for our experiments. We then show how to extend their feature set and rely on active learning to improve the classification. As mentioned above, the dataset provided by Fernández, Ginzburg, and Lappin (2007) is composed of 1283 manually annotated non-sentential utterances. Each instance contains the identifier of the NSU in the BNC, the identifier of its antecedent, and the class of the NSU selected by the annotator. Following the same method as Fernández, Ginzburg, and Lappin (2007), we also filter the dataset by considering only NSUs whose antecedent is the preceding utterance. This facilitates the feature extraction task, and only leads to a minor reduction of the number of instances (12% of the total), as the vast majority of the annotated NSUs have their immediately preceding utterance as antecedent. After this filtering operation, the dataset contains a total of 1123 instances.

3.1 Baseline features

The approach of Fernández, Ginzburg, and Lappin (2007) relied on a feature set composed of 9 features divided into three groups: *NSU features*, *antecedent features*, and *similarity features*.

The first group contains four features that exploit some key syntactic and lexical patterns in the surface form of the NSU, such as the mood of the NSU (proposition or question), the occurrence of *wh*-words (*who*, *where*, *how*, . . .), the presence of affirmative (*yes*, *yeah*, *yep*, . . .), negative (*no*, *not*, *nope*, . . .) or acknowledgement words (*mhm*, *okay*, *right*, . . .), or the presence of modal adverbs or adjectives (*good*, *great*, *probably*, . . .) at the onset of the NSU. These features are useful to recognize interrogative NSUs such as sluices and clarification ellipsis, as well as NSUs used to modify or extend the antecedent, such as propositional modifiers and conjuncts.

Three antecedent features indicate specific properties of the antecedent of an NSU. They encode whether the antecedent is declarative or not, whether it contains *wh*-words, and whether it is truncated at the end or not. These features can help assessing whether, for instance, the antecedent is a *wh*-question, in which case the NSU is likely to be a short answer or a request for clarification/information. The combination of NSU and antecedent features can help predict the class of most of the NSUs.

The last group, the similarity features, includes two features that exploit, at the syntactic level, the parallelism between the NSU and its antecedent. One feature counts the number of content words in common, while the second encodes the number of common sequences of POS tags (from the BNC tag set) in the antecedent and NSU. These features seek to capture some typical patterns of repeated acknowledgments, repeated affirmative answers and helpful rejections.[1]

3.2 Extended features

To improve the classification performance beyond the baseline, the feature set was extended with 23 novel features. These features were added to the baseline feature set, for a total of 32 features in the final feature set. The new features have been developed in

[1] For more detail on how the baseline feature set was replicated, refer to Dragone (2015).

order to incorporate more detailed information about the linguistic structure of the NSU and its local conversational context. These features exploit deeper syntactic and lexical patterns such as syntactic tags and dependency relations, as well as dialogue-level features and additional similarity measures between NSUs and their antecedents. The additional features can be divided into five groups: *surface features*, *phrase-level features*, *dependency features*, *turn-taking features*, and *similarity features*.

The first group is extracted using the annotation already available in the BNC. The group contains 7 features, of which 4 are the POS tags of the first four words in the NSU; the remaining 3 encode the presence of an ending punctuation, a pause or an unclear word or phrase in the antecedent.

The phrase-level and the dependency-level features indicate the presence of syntactic patterns in the NSU and the antecedent. The phrase-level group contains 7 features extracted by parsing the antecedent with the PCFG parser of the Stanford CoreNLP library (Klein and Manning 2003). Three of these features signal the presence of the `SQ` (inverted yes/no question), `SBARQ` (direct question introduced by a wh-word), and `SINV` (inverted declarative sentence) syntactic tags in the antecedent. These syntactic tags are useful to recognize questions, especially when an explicit question mark is missing. Three other features indicate the first clause-level tag, phrase-level tag, and POS tag (from the PCFG parser tag set) of the NSU; the last phrase-level feature indicates the presence of a pattern for the *negation + correction* in the NSUs, as in the following example:

A: Or, or were they different in your childhood?
B: **No, always the same.** [BNC: JK8 137-138]

The group of dependency features comprises two features that signal the presence of patterns in the dependency relations of the antecedent, as generated by the dependency parser of the Standord CoreNLP (Chen and Manning 2014). The first looks for *neg* dependencies (generally arising from adverbial negations such as *not*, *don't*, *never*), which can be used to capture an affirmative answer expressed in a negated form, as in this example:

A: You're not getting any funny fits from that at all, June?
B: **Er no.** [BNC: H4P 36–37]

The second dependency feature looks for *wh-interrogative* fragments, regardless of the presence of a question mark (details in Dragone (2015)).

The turn-taking features currently contains only one feature indicating whether or not the NSU and its antecedent were uttered by the same speaker. This is typical of check questions that request an acknowledgement of understanding from the listener(s).

The similarity feature group adds six new features to the other two in the baseline feature set. Despite its name, this group also contains features that are not, strictly speaking, similarity measures, namely the number of words in the NSU (both all words and only content words). Other features in this group measure multiple types of similarity at character, word and POS-level between the NSU and its antecedent. Two of these features are extracted using a modified version of the Needleman-Wunsch algorithm to determine the longest common subsequence of words and POS tags (Needleman and Wunsch 1970). Another feature uses the Smith-Waterman algorithm for the local alignment at the character level (Smith and Waterman 1981). Finally, the last similarity feature encodes the number of words contained in the last part of the antecedent that are also observed in the non-sentential utterance.

3.3 Active learning

The dataset employed for this classification task has a number of important shortcomings. The first one is the scarcity of annotated data, as the complete dataset only comprises 1123 instances. In addition, the available annotated data suffers from class imbalance, as evidenced in Table 1. The five most common NSU classes indeed constitute more than 82 % of the annotated data, while the five least common classes constitute less than 6 % of the same data. The combination of these two problems makes the classification task quite difficult, especially since the low-frequency classes are also typically the most difficult to predict. On the other hand, large amounts of unannotated data is available from the BNC (or similar types of spoken language corpora) and can be extracted based on simple heuristics, making additional annotation appealing for reducing the impact of the aforementioned problems. However, the fact that the low-frequency classes are typically the ones that are hard to predict means that a large number of annotations would need to be collected in order to improve the classification accuracy. For this reason, we employed *active learning* to address the two related issues of data scarcity and class imbalance. Active learning (Settles 2010) is a machine learning technique used to enrich the available data with new instances by querying an expert user for the label of some unlabeled instances. The active learning algorithm selects the most "informative" instance among a pool of unlabeled instances.

NSU detection

The first step in exploiting the unlabelled data in the BNC is to detect the NSUs, i.e. decide whether a given utterance is an non-sentential utterance or not. This can be achieved using straightforward heuristics. In particular we defined a set of rules to determine whether an utterance is an NSU or not (see Dragone (2015) for more details in the detection procedure). The most important rules are the following:

- NSUs are short, i.e. the number of words must be less then a threshold;
- Greetings and other short sentences that are not NSUs are discarded;
- The sentence must not contain a verb phrase in any form.

Using these rules on the 1123 NSU instances in the labeled dataset, we correctly detected 1033 instances, with an accuracy of 92%. We restricted the extraction of potential NSUs to feed in the active learning process to two-party dialogues. The motivation behind this restriction is to facilitate the selection of the NSU antecedents. Indeed, Fernández, Ginzburg, and Lappin (2007) shows that in two-party dialogues the percentage of NSUs whose antecedent is not their preceding utterance is about 7%, compared to 19% in multi-party dialogues. We also constrain the antecedent (assumed to be the preceding utterance) to be longer than the NSU and to have a complete clausal form, i.e. with at least a noun phrase and a verb phrase. Using these heuristics on the BNC dialogues, we extracted in total 3198 unlabeled NSU instances.

Selection of instances to query

Among the many strategies for selecting an informative instance, we employed an "uncertainty sampling" method using *entropy* as uncertainty measure (Lewis and Catlett

1994). Using the available data, we train a classifier to predict the probability distribution of the classes for all the unlabeled instances. For each unlabeled instance i, we compute the entropy over the distribution of the classes C given the feature vector $\mathbf{f}_i$. We then select the instance with the highest entropy:

$$i^* = \text{argmax}_i \; H(C_i) = \text{argmax}_i \; \left(- \sum_{c \in C} P(C_i = c | \mathbf{f}_i) \log P(C_i = c | \mathbf{f}_i) \right) \tag{1}$$

This process is then repeated, training a new classifier after adding the newly labeled instance to the labeled pool, until the goal is reached. The entropy measures the "unpredictability" of the class of an instance. The higher the entropy, the more difficult is to discriminate the right class from the others. Knowing the class of the unlabeled instance with highest entropy gives the most information. Compared to other strategies, entropy sampling is particularly helpful for multinomial classification tasks (Settles 2010). For our final approach we annotated 100 new NSU instances via this active learning process. Table 2 shows the class frequency of the new instances. The active learning algorithm selects instances of "difficult" classes, which are also the low frequency ones. This has the advantage to add more information at a lower annotation cost.

Table 2
Class frequencies of the 100 additional NSUs extracted via active learning.

NSU Class	Instances
Helpful Rejection	21
Repeated Acknowledgment	17
Clarification Ellipsis	17
Acknowledgment	11
Propositional Modifier	9
Filler	9
Sluice	3
Repeated Affirmative Answer	3
Factual Modifier	3
Conjunct Fragment	3
Short Answer	2
Check Question	2

4. Evaluation

This section describes the empirical evaluation of our classification approach on the annotated corpus. We first detail the baseline employed for the evaluation, and then compare its results with the ones obtained with the approach combining the extended feature set with the active learning strategy.

4.1 Baseline replication

To allow for a detailed comparative analysis of the evaluation results, we first replicated the classification experiments detailed in Fernández, Ginzburg, and Lappin (2007), which we use as baseline. The classifier is estimated with the Weka J48 algorithm for decision trees learning, the same used by Fernández, Ginzburg, and Lappin (2007). The performance scores of our replica are comparable with the ones presented in their article, apart from the accuracy which was not provided in the original article. Unfortunately, we could not make a fine-grained statistical comparison since we did not have access to the original feature extraction algorithm. Table 3 shows the comparison scores between our replica and the scores presented in Fernández, Ginzburg, and Lappin (2007).

Table 3
Performances comparison between Fernández, Ginzburg, and Lappin (2007) and our replica.

	Replicated experiment			**Reference classification**		
NSU Class	**Precision**	**Recall**	**F_1-score**	**Precision**	**Recall**	**F_1-score**
Plain Acknowledgment	0.97	0.97	0.97	0.97	0.95	0.96
Affirmative Answer	0.89	0.84	0.86	0.83	0.86	0.84
Bare Modifier Phrase	0.63	0.65	0.62	1.00	0.70	0.82
Clarification Ellipsis	0.87	0.89	0.87	0.92	0.92	0.94
Check Question	0.85	0.90	0.87	0.83	1.00	0.91
Conjunct Fragment	0.80	0.80	0.80	0.71	1.00	0.83
Factual Modifier	1.00	1.00	1.00	1.00	0.91	0.95
Filler	0.77	0.70	0.71	0.50	0.37	0.43
Helpful Rejection	0.13	0.14	0.14	0.46	0.33	0.39
Propositional Modifier	0.92	0.97	0.93	1.00	0.60	0.75
Rejection	0.76	0.95	0.83	0.76	1.00	0.86
Repeated Acknowledg.	0.74	0.75	0.70	0.84	0.86	0.85
Repeated Aff. Answer	0.67	0.71	0.68	0.65	0.68	0.67
Short Answer	0.86	0.80	0.81	0.81	0.83	0.82
Sluice	0.67	0.77	0.71	0.95	1.00	0.98
Micro-Average	**0.89**	**0.89**	**0.88**	**0.90**	**0.90**	**0.89**

4.2 Results with extended features and active learning

Our experiments were carried out using Weka, a Java package for machine learning (Hall et al. 2009)[2], and JCLAL, a Java library for active learning[3]. We first collect the aforementioned 100 newly labeled instances by running the active learning algorithm. For the active learning procedure we split the dataset into three partitions (50-25-25 %). We use the training set (50%) as input of the Weka's SMO classifier, an SVM implementation using the sequential minimal optimization algorithm. The SMO classifier is used by the active learning process to predict the classes of unlabeled instances. The development set (25%) is used to evaluate the classifier as instances are added to the training set. The test set (25%) is finally used to compare the various settings at the end of the experimental phase. From this evaluation we can see that the addition of new instances is only beneficial when using the extended feature set (see Table 4).

[2] cf. `http://www.cs.waikato.ac.nz/ml/weka/`.
[3] cf. `http://sourceforge.net/projects/jclal`.

Table 4
Performances of the SMO classifier in the various settings on the test set

Training set (feature set)	Accuracy	Precision	Recall	F_1-score
Train-set + Dev-set (baseline)	0.906	0.911	0.906	0.903
Train-set + Dev-set (extended)	0.928	0.937	0.929	0.930
Train-set + Dev-set + active learn. (basel.)	0.898	0.911	0.898	0.898
Train-set + Dev-set + active learn. (extend.)	**0.932**	**0.945**	**0.932**	**0.935**

Table 5
Accuracy, precision, recall and F_1 scores for each experiment, based on the J48 classifier, with 10-fold cross validation.

Training set (feature set)	Accuracy	Precision	Recall	F_1-Score
Train-set (baseline)	0.885	0.888	0.885	0.879
Train-set (extended)	0.889	0.904	0.889	0.889
Train-set + active learning (baseline)	0.890	0.896	0.890	0.885
Train-set + active learning (extended)	0.896	0.914	0.896	0.897

Table 6
Accuracy, precision, recall and F_1 scores for each experiment, based on the SMO classifier, with 10-fold cross validation.

Training set (feature set)	Accuracy	Precision	Recall	F_1-Score
Train-set (baseline)	0.881	0.884	0.881	0.875
Train-set (extended)	0.899	0.904	0.899	0.896
Train-set + active learning (baseline)	0.883	0.893	0.883	0.880
Train-set + active learning (extended)	**0.907**	**0.913**	**0.907**	**0.905**

A further evaluation is performed using 10-fold cross validation over the full dataset, using the SMO algorithm and the Weka's J48 algorithm, an implementation of the C4.5 algorithm for decision trees (Quinlan 1993), also used by Fernández, Ginzburg, and Lappin (2007). We compared the results for every setting using both algorithms. Table 5 shows the performance achieved by the J48 algorithm whereas Table 6 shows the results for the SMO classifier. The addition of the new instances did not significantly improve the performance of the J48 classifier. However, we can observe a statistically significant improvement in the accuracy performance of the SMO algorithm between the baseline and the final approach (using the extended feature set and active learning), based on a pairwise t-test with a 95% confidence interval, for a p-value $= 6.9 \times 10^{-3}$.

The SMO algorithm does not initially provide the best accuracy for the baseline, but the use of extended features does seem to improve its classification performance better than the J48 algorithm. This evaluation confirms that to achieve a significant improvement both the extended feature set and the additional training data are needed.

Table 7 shows the detailed per-class results for the baseline and the final approach. While there is an improvement for many classes, some classes remain nevertheless difficult to predict, such as helpful rejections and repeated fragments. Helpful rejections are particularly hard to classify because the parallelism with their antecedent is mostly semantic, as in the following example:

A: There was one which you said Ernest Morris was born in 1950.
B: **Fifteen.** [BNC: J9A 372-373]

To classify this kind of non-sentential utterance, syntactic and lexical features are indeed not sufficient, as "fifteen" is intended here as a correction of the year mentioned in the antecedent (the year was 1915, not 1950). Deeper semantic analysis is required to extract features able to capture these kind patterns.

Table 7
Precision, recall and F_1 score per class between the baseline (initial feature set and J48 classifier) and the final approach (extended feature set with active learning and SMO classifier).

	Baseline			**Final approach**		
NSU Class	**Precision**	**Recall**	**F_1-Score**	**Precision**	**Recall**	**F_1-Score**
Plain Ack.	0.97	0.97	0.97	0.97	0.98	0.97
Affirmative Answer	0.89	0.84	0.86	0.81	0.90	0.85
Bare Modifier Phrase	0.63	0.65	0.64	0.77	0.75	0.76
Clarification Ellipsis	0.87	0.89	0.87	0.88	0.92	0.90
Check Question	0.85	0.90	0.87	1.00	1.00	1.00
Conjunct Fragment	0.80	0.80	0.80	1.00	1.00	1.00
Factual Modifier	1.00	1.00	1.00	1.00	1.00	1.00
Filler	0.77	0.70	0.73	0.82	0.83	0.82
Helpful Rejection	0.13	0.14	0.13	0.31	0.43	0.36
Propositional Modifier	0.92	0.97	0.94	0.92	1.00	0.96
Rejection	0.76	0.95	0.84	0.90	0.90	0.90
Repeated Ack.	0.74	0.75	0.74	0.77	0.77	0.77
Repeated Aff. Answer	0.67	0.71	0.69	0.72	0.55	0.62
Short Answer	0.86	0.80	0.83	0.92	0.86	0.89
Sluice	0.67	0.77	0.71	0.80	0.84	0.82

5. NSU resolution

In this section, we briefly describe a novel approach to the *resolution* of NSUs. Due to space constraints, here we will address this issue on a fairly abstract way, referring to Dragone (2015) for more details.

The resolution, as described by e.g. Fernández (2006), is the second step towards the interpretation of the meaning of an NSU. This task builds up on the classification by exploiting the information given by the predicted type of an NSU. The goal of the resolution process is to determine how the non-sentential utterance precisely updates the current state of the dialogue, by e.g. introducing new information, raising new questions or providing positive or negative feedback on each other's contributions to the dialogue.

Logical frameworks

Traditional approaches to NSU resolution are based on logical *resolution rules*. For each NSU class, Fernández (2006) defines one specific resolution rule. Given the raw NSU and its type as input, the objective of the rules is to extract a semantic representation of such NSU from the *dialogue context*. The dialogue context is generally represented as a set of variables that keeps track of the information state throughout the evolution of the dialogue.

One standard method to represent such a dialogue context is to employ a logical formalism such as Type Theory with Records (TTR) (Cooper 2005). This formalism is notably employed in the formal theory of conversation developed by Ginzburg (2012). In this theory, the dialogue context is represented in terms of TTR records encompassing elements such as the set of known facts (propositions assumed to be true by all conversational partners), the last dialogue act(s) in the conversation history, and an ordered set of *questions under discussion* to resolve in the course of the interaction. The evolution of the conversation is then formalised by means of update rules operating on this dialogue context.

One limitation of these logical frameworks is their inability to directly represent (and reason over) uncertain knowledge. Many aspects of the dialogue context are indeed only partially observable, such as the actual user intentions or the set of perceived entities in the visual scene. Computational models of dialogue must also account for a certain degree of stochastic behaviour in the conversation dynamics, as the evolution of a given dialogue is not a deterministic process. The stochastic component is especially important in dealing with NSUs since they do not have a full-fledged meaning isolated from their context and are, as argued in Ginzburg (2012), in principle highly ambiguous.

Probabilistic rules

To address these issues, we developed a new approach to the resolution of NSUs that relies on probabilistic modelling to encode and update the dialogue context. More specifically, we employ the probabilistic rules formalism of Lison (2015) to encode the NSU resolution procedure.

Probabilistic rules are similar, to a certain extent, to the update rules of Ginzburg (2012), but they can be applied on dialogue states containing partially observable variables. Furthermore, they can include unknown parameters that can be automatically estimated from dialogue data. Probabilistic rules are technically defined as *if...then...else* constructions associating logical conditions on input variables to probabilistic effects over output variables. These rules function as high-level templates for the generation of a directed graphical model (see Lison (2014, 2015) for details). Their support for logical abstraction is particularly useful to encode the (sometimes complex) update semantics of non-sentential utterances.

We employed the OpenDial toolkit (Lison and Kennington 2016) to develop a proof-of-concept implementation of this approach. The dialogue state is represented as a Bayesian Network including various aspects of the interaction such as the questions under discussion, the set of currently known facts and the last dialogue acts from every dialogue participant. A small set of probabilistic rules is then specified to encode the resolution procedure for each NSU type, following the formalisation developed by Ginzburg (2012). For instance, a short answer "Robert" following the question "What is your name?" triggers the update of the dialogue state by introducing a new fact, namely that the current speaker is named Robert. Similarly, a sluice will typically introduce a new question under discussion, with a semantic content that depends both on the

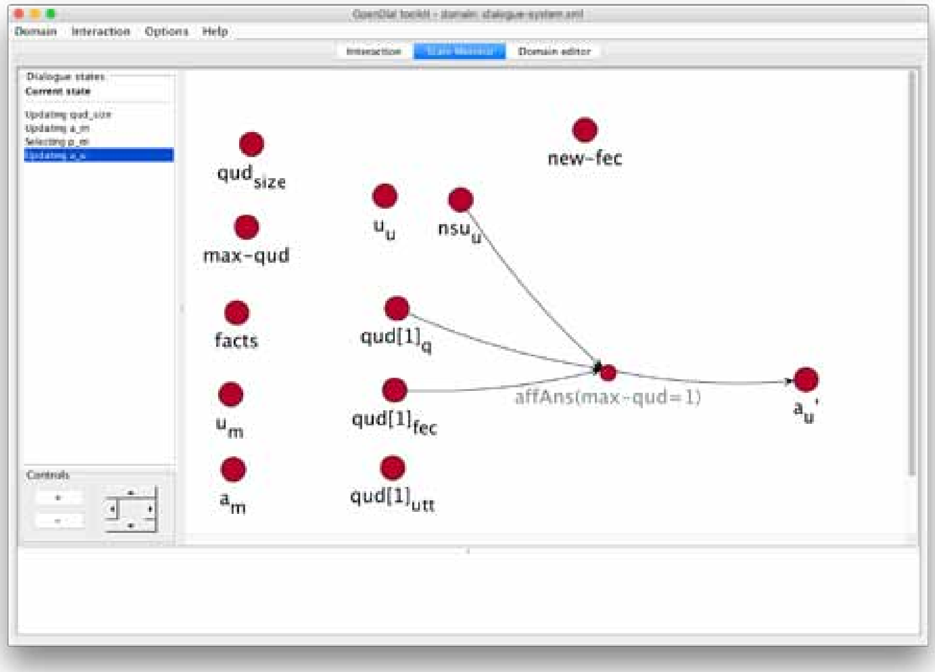

Figure 1
Example of state update in OpenDial following the occurrence of a new non-sentential utterance in the course of the dialogue. The dialogue state is factored into several variables, each being represented by a specific node of the Bayesian Network. The technical description of each individual variable in this example is beyond the scope of this article, but we can observe that the state contains multiple variables related to the current questions under discussion (QUD), focus-establishing constituents (FEC), user and system dialogue acts (a_u and a_m) along with their corresponding raw utterances (u_u and u_m). The nsu_u variable denotes the predicted class of the non-sentential utterance, as provided by the classifier. The update step in this screenshot shows the application of a probabilistic rule updating the value of the user dialogue act a_u following the detection of a new non-sentential utterance (here an affirmative answer). We can observe that the value of the new user dialogue act is dependent on both the detected NSU class and the state variables encoding the current question under discussion. Technical details about the state variables and the probabilistic rules operating on them can be found in Dragone (2015).

sluice being used and the constituents present in the antecedent. Figure 1 provides an illustration of the update process resulting from the application of a probabilistic rule on the dialogue state.

The probabilities employed in the rules developed for this proof-of-concept implementation were handcrafted, but can in principle be estimated from actual dialogue data, as explained in (Lison 2015). Detailed descriptions of probabilistic rules for NSU resolution along with examples of application on short dialogues can be found in (Dragone 2015).

6. Discussion

Most of the existing literature on non-sentential utterances originates from the field of formal semantics (Dalrymple, Shieber, and Pereira 1991; Fernández 2006; Schlangen 2003; Ginzburg 2012). There has been by comparison fewer studies of non-sentential utterances from an NLP perspective.

The classification approach detailed in Section 3 built upon the taxonomy of Fernández, Ginzburg, and Lappin (2007). It should be noted that other taxonomies of non-sentential utterances have been proposed, such as the coherence-based approach developed by Schlangen (2003). This approach, however, rests on a particular theory of discourse, namely *Segmented Discourse Representation Theory* (Asher and Lascarides 2005), which makes it more difficult to apply outside of this particular theoretical framework. A detailed comparison of this taxonomy and other ones can be found in (Fernández 2006), who also details the corpus study on the BNC corpus that led to the definition of this taxonomy.

To a certain extent, the task of classifying NSUs is complementary to the correct identification of such utterances. This problem has been tackled in the past by Schlangen (2005), which focuses on the detection of NSUs in multi-party dialogues. His work provides an interesting approach useful not only for the detection of the NSUs but also for the identification of the antecedents. For the extraction of unlabeled data from the BNC, which we described in Section 3.3, we did not need to address this problem directly, since we limited our scope to two-party dialogues only. The NSU identification method of Schlangen (2005) would, however, constitute a promising strategy to extend the proposed approach to generic, multi-party dialogues.

As argued by Raghu et al. (2015), the taxonomy from Fernández, Ginzburg, and Lappin (2007) does not cover all possible NSU classes, possibly because their study was concentrated on a specific collection of dialogues from the BNC corpus. Their own work concentrates on a type of NSUs identified by follow-up questions, which they argue do not fit well with the taxonomy of Fernández, Ginzburg, and Lappin (2007).

The empirical results presented in this article also highlight the need for a larger annotated corpus of non-sentential utterances. In our view, the development of such a corpus, including new dialogue domains and a broader range of conversational phenomena, could contribute to a better understanding of non-sentential utterances and their semantic interpretation.

The NSU resolution problem has been addressed by several previous approaches. The system developed by Purver (2006) handles clarification requests, including the ones with elliptical form. Ginzburg et al. (2007) targets the full range of NSU classes, based on the theoretical framework developed by Fernández (2006). These approaches originate from the theoretical framework of Ginzburg and Sag (2000), which is then further developed by Ginzburg (2012). The main difference between these approaches and the one presented in Section 5 is the fact that the latter encodes the dialogue state as a partially observable representation rather than a logical construct. The work of Raghu et al. (2015) is also related to this work. Their approach relies on a purely statistical model used to rank the possible reconstructed sentences starting from the keywords in the antecedent and the follow-up questions. While this approach is fully data-driven, it lacks a semantic representation of the sentences, and is therefore difficult to generalise to other kinds of NSUs. Our approach can be seen as adopting a hybrid logical/probabilistic strategy, taking advantages of the fine-grained semantic framework of Ginzburg (2012) while extending it to allow for a probabilistic treatment of ambiguities and partially observable contexts.

7. Conclusion

This paper presented a novel approach to the classification of non-sentential utterances, extending the work of (Fernández, Ginzburg, and Lappin 2007). The approach relied on an extended feature set and active learning techniques to address the scarcity of labelled data and the class imbalance. The evaluation results demonstrated a small but statistically significant improvement in the classification accuracy.

The last part of this paper also sketched a probabilistic account of the resolution of non-sentential utterances based on its surrounding dialogue context. This approach relies on a small number of probabilistic rules to specify the update semantics associated with various types of non-sentential utterances, based on a dialogue state represented as a Bayesian Network.

The presented approach can be extended in multiple directions, such as the use of semantic or discourse-level features to improve the classification of non-sentential utterances. Another interesting venue for future work is to investigate how to automatically estimate the parameters of the probabilistic NSU resolution rules from actual dialogue transcripts.

References

Asher, Nicholas and Alex Lascarides. 2005. *Logics of Conversation*. Cambridge University Press.

Burnard, Lou. 2000. Reference guide for the british national corpus (world edition).

Chen, Danqi and Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 1, pages 740–750.

Cooper, Robin. 2005. Records and record types in semantic theory. *Journal of Logic and Computation*, 15(2):99–112.

Dalrymple, Mary, Stuart M. Shieber, and Fernando C. N. Pereira. 1991. Ellipsis and higher-order unification. *Linguistics and philosophy*, 14(4):399–452.

Dragone, Paolo. 2015. Non-sentential utterances: Experiments in classification and interpretation. Master's thesis, Sapienza University of Rome.

Fernández, Raquel. 2006. *Non-Sentential Utterances in Dialogue: Classification, Resolution and Use*. Ph.D. thesis, King's College London.

Fernández, Raquel, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying non-sentential utterances in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427.

Ginzburg, Jonathan. 2012. *The Interactive Stance*. Oxford University Press.

Ginzburg, Jonathan, Raquel Fernández, Howard Gregory, and Shalom Lappin. 2007. SHARDS: Fragment resolution in dialogue. *Computing Meaning*, 4:125–144.

Ginzburg, Jonathan and Ivan Sag. 2000. *Interrogative investigations*. Stanford: CSLI publications.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Lewis, David D. and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the eleventh international conference on machine learning*, pages 148–156.

Lison, Pierre. 2014. *Structured Probabilistic Modelling for Dialogue Management*. Ph.D. thesis, University of Oslo.

Lison, Pierre. 2015. A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech & Language*, 34(1):232 – 255.

Lison, Pierre and Casey Kennington. 2016. OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (System Demonstrations)*, Berlin, Germany. Association for Computational Linguistics.

Needleman, Saul B. and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Purver, Matthew. 2006. CLARIE: Handling clarification requests in dialogue system. *Research on Language and Computation*, 4(2-3):259–288.

Quinlan, Ross J. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.

Raghu, Dinesh, Sathish Indurthi, Jitendra Ajmera, and Sachindra Joshi. 2015. A statistical approach for non-sentential utterance resolution for interactive qa system. In *Proceedings of the SIGDIAL 2015 Conference*, pages 335–343.

Schlangen, David. 2003. *A coherence-based approach to the interpretation of non-sentential utterances in dialogue*. Ph.D. thesis, University of Edinburgh. College of Science and Engineering. School of Informatics.

Schlangen, David. 2005. Towards finding and fixing fragments: Using ML to identify non-sentential utterances and their antecedents in multi-party dialogue. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 247–254. Association for Computational Linguistics.

Settles, Burr. 2010. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11.

Smith, Temple F. and Michael S. Waterman. 1981. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197.