# TalKit - Desenvolvimento de um Sistema de Diálogo para Português

Cátia Dias
catia.dias@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

May 2015

**Abstract**

In this thesis we propose taxonomy for classifying interactions given by a user to a dialogue system and developed TalKit, with the purpose of responding to different types of interactions, choosing the modules responsible for the response based on that taxonomy. Two existing modules, the Say Something Smart and Talkpedia were used in this process, having the second been extended in this work. Multiple classifiers were implemented, which combine handwritten rules and modules created with machine learning techniques, to map the interactions made by a user, in taxonomy developed classes. Two strategies were tested for the implementation of TalKit. The first one based on machine learning and rule-based. The second strategy based, again, on machine learning, having been created the `Mega Classificador`, which uses features the rules previously developed, as well as the rest of the classifiers. After the evaluation of these two strategies, it was concluded that the first is better than the second, since the first manages to classify correctly 87% interactions and the second 80%. Finally, the classifier based on machine learning and on rule-based was used to decide which module to choose to obtain a response and was used in an evaluation done with users. It was concluded that using the TalKit obtains more plausible answers than using just the Talkpedia or the Say Something Smart. Additionally it was proven that the extensions made to the Talkpedia, increase the number of plausible responses returned by the system.

**Keywords:** Natural Language Understanding, Dialogue Taxonomy, Classification, Machine Learning, Rule-Based

## 1 Introduction

Currently it has been registered the emergence of virtual assistants, that help us with specific tasks, such as *Siri* from Apple[1], *Anna* from IKEA or *Edgar Smith* [8], the butler of Monserrate's Palace, in Sintra. Even though these systems are oriented to a particular domain, such as butler *Edgar* which answers questions about the Palace, users tend to question them with questions outside of their domain. So to be able to answer these questions, it is necessary to develop systems that can automate the process of giving an answer that's not present on their base of information [12]. Systems that handle these type of situations were already developed by L2F[2], i.e. that deal with out-of-domain interactions, namely:

- The Talkpedia [12], which is a module that responds to out-of-domain questions, based on the *Wikipedia*[3].
- The Say Something Smart [2], which is a system that responds based on movie subtitles.

Each one of these systems (isolated) can respond to a specific set of questions. For example, the Say Something Smart is directed to respond to personal interactions (example: *How old are you?*); now the Talkpedia, answers to factoids questions (example: *Where is Lisbon?*). Therefore, it makes sense to have a system that can distinguish the various types of interactions, given by a user, in order to redirect the interaction to one of these systems and/or other information bases.

In this thesis was developed the TalKit. The TalKit is a dialog system that responds to interactions, where in its knowledge is drawn from various sources (answer search module). For the TalKit to know where to get the answer, taxonomy was developed to classify the dialog interactions. Accord-

[1] https://www.apple.com/ios/siri/
[2] https://www.l2f.inesc-id.pt
[3] https://www.wikipedia.org/

ing to the classification given to the interaction, the system redirects it to the most appropriate answer search module.

## 2 Background

There are several projects previously developed by L2F which are connected to this work. As the Talkpedia system [12], which has the job of handling out-of-domain interactions in order to give a coherent answer to the user, thus making the interaction more appealing, interesting and less monotonous. The Talkpedia knowledge base is the online encyclopedia Wikipedia. Another system is the Say Something Smart [2]. It is a system that indexes a corpus and selects the most appropriate answer for a question, carried out by a user. The Just.Ask [15] is a Question-Answer (QA) system which combines rule-based components with machine learning. The Just.Chat [14] is a platform that has as main objective the development of knowledge bases, from chatbots interactions, to later be used by other chatbots.

## 3 Related Work

Over the years, several dialog taxonomies were developed in order to classify the interactions performed by a user.

The Dialog Act Markup in Several Layers (DAMSL) [6, 1] is a taxonomy focused on task-oriented dialogues. The taxonomy consists of four main categories: Communicative Status, Information Level, Forward Looking Functions and Backward Looking Functions. The Communicative Status category indicates whether a phrase expressed by a user is intelligible and if it was completed successfully. In the Information Level category the semantic content of an expression is characterized. The Forward Looking Functions category indicates the effect of an expression in a given dialogue, i.e. which action the expression expresses in a dialogue. An example of this category can be given by the following phrase *"please take out the trash"*. Finally, the category Backward Looking Functions refers to whether the contents of the current declaration belong to a previous dialogue.

The taxonomy Meeting Recorder Dialog Act (MRDA) [7] is an adaptation of the SWBD-DAMSL [10] taxonomy, and consists of 13 groups, where each group contains a set of tags.

The Dynamic Interpretation Theory (DIT) [4] is a computational approach that analyzes the meaning of expressions in a dialogue between humans or an interaction between man and computer. Over the years the DIT has been extended and modified, taking into account other proposed taxonomies (such as DAMSL [6], MRDA [7], among others), which resulted in the taxonomy DIT++[4] [5, 9]. The DIT++ [9] consists of two parts: the General-purpose Functions (functions that are applied on any type of semantic content) and the Dimension-specific Functions.

The Dialogue Scheme for Unifying Speech and Semantic (DISCUSS) [3] is a dialogue move taxonomy with the purpose of seizing semantic and pragmatic interpretations of expressions. It consists of two dimensions: the Semantic Dimension (which consists of the dimensions Predicate Type and Semantic Roles) and the Pragmatic Dimension (which is composed of Dialogue Act and Rhetorical Form).

Beyond dialog taxonomies, there are also taxonomies to solely classify questions that are typically used by question-answer systems. One of the best-known taxonomies is the Li and Roth's taxonomy [11]. It's a hierarchical taxonomy, consisting of two layers. The first layer (the most in-depth, by the name *Coarse*) contains six categories (ABBREVIATION, DESCRIPTION, ENTITY, HUMAN, LOCATION and NUMERIC). The second layer contains fifty *Fine* categories (that is, contains 50 specified categories), where each of them corresponds to one of the classes of the first layer. For example, the classification of the question *"Which countries border Spain?"* is LOCATION: COUNTRY.

## 4 Taxonomy Proposal

When a dialog line enters the system is classified according to its type, and can be of type question or non-question, personal or impersonal type, among others.

The proposed taxonomy took into account the various taxonomies referred to earlier (Section 3). So, to start, the task of classification is based on distinguishing if a dialog line is a question or not, which may have one of the following tags:

- QUESTION - The interaction given by a user is a question (example: *Que dia da semana é hoje?*).
- NON_QUESTION – The interaction given by a user is a non-question (example: *Hoje está muito calor.* or *Bom dia!*).

If a dialog line is classified as QUESTION can additionally be sub-classified as follows:

- YN_QUESTION - Questions with answers like "yes" or "no" (example: *Tens um gato?*).

[4] http://dit.uvt.nl/

- OR_QUESTION – Option questions (i.e. which contains the conjunction "ou") (example: *Tens um cão ou um gato?*).
- RHETORICAL_QUESTION – Rhetorical questions, i.e., questions in which is not normally expected a reply (example: *A sério?*).
- OPEN_ENDED_QUESTION – Open answer questions, i.e., questions that do not have a specific answer (example: *mais alguma coisa?*).
- LIST_QUESTION - Questions that express a request to indicate something (may be a set of names, just a name, an indication, etc.). They are normally initiated by the conjugated verbs "Indicar", "Dizer", "Referir", among others. This type of questions does not contain the "¿" (example: *Indique dois jogos de tabuleiros.*).
- WH_QUESTION - Questions that begin with an interrogative adverb ("quando", "porque", "como", "onde") or na interrogative pronoun ("qual", "quanto", "quem", "que") (example: *Quem foi Amália Rodrigues?* ou *Que idade tens?*).

The classified questions with one of the tags mentioned previously can be further sub-classified as personal or impersonal questions by using the following tags:

- PERSONAL - The question given by a user is of type personal (example: *Como te chamas?*).
- IMPERSONAL - The question given by a user is of type impersonal (exemplo: *Quem é Alain Juppé?*).

If the question was classified by the IMPERSONAL label, this can be further classified by the taxonomy proposed by Li & Roth [11].

If, however, the dialog line was classified as not being a question (NON_QUESTION), this can be sub-classified with one of the following tags:

- SOCIAL_OBLIGATIONS - Dialog lines expressing greetings or goodbyes (example: "Olá!", "Bom dia", "Obrigado", "Desculpa", "Adeus", among others).
- ACKNOWLEDGEMENTS - Dialog lines that are shaped like confirmation and normally do not have feedback (example: *A: "Lisboa é a capital de Portugal" U: "Ok!"* wherein A refers to the referred agent and the user U).
- DECLARATIVE_SENTENCE - Declarative type dialog lines (example: In the previous case the sentence "Lisboa é a capital de Portugal" is a DECLARATIVE_SENTENCE).

In addition to the labels referred to earlier, a dialog line may also be classified with the tag NO_UNDERSTANDING. This tag classifies phrases/expressions that indicate that the user didn't notice or was confused by the answer given by the system (example: *A: "Porto é um vinho." U: "Não percebi" or U: "Podes repetir?"*). These expressions can be and/or contain interjections such as "ah", "hum", among others.

# 5 TalKit Architecture

TalKit's architecture, such as shown in the Figure 1, consists of a pipeline composed of three main modules: `Pré-processamento`, `Classificador` and `Procura de Resposta`. Briefly, the TalKit receives as input a dialog line (which may be a question or not), which passes by the `Pré-processamento` module where the dialog line is analyzed in order to be taken off a set of information. Next, the dialog line pre-processed is forwarded to the `Classificador` module where it will be sorted according to the proposed taxonomy. Based on this classification, is directed to the appropriate module (module of `Procura de Resposta`), which returns a answer.

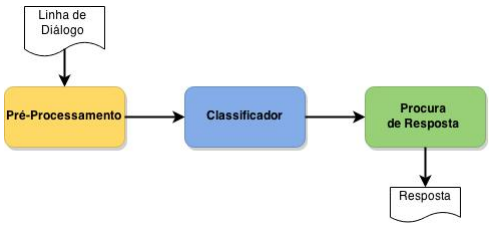


Figure 1: TalKit architecture.

# 6 Implementation

This section presents the corpora used, as well as the implementation of rule-based classifier and machine learning classifier. Were also explained two strategies proposed for the classification of sentences. This section also describes the extensions made to the Talkpedia.

## 6.1 Rule-Based Classifier

### 6.1.1 Corpora

It was created a corpus with 360 questions, in which 180 are yes/no questions and other questions belong to other categories. And was also created a corpus in which contains 18 rhetorical questions, 6 open answer questions, 19 relating to expressions

of greetings and goodbyes, 13 expressions that indicate a confirmation and 16 questions/phrases that indicate that the user does not understand what was said.

#### 6.1.2 Rules

The classification of rules-based dialog lines consists of assigning a tag to them if these correspond to a set of rules. Therefore, several patterns have been developed (rules based on regular expressions) for the following tags:

- **Tag YN_QUESTION**
  To identify yes/no questions the classifier uses a set of rules implemented manually and based on the TreeTagger[5] tool.

- **Tag OR_QUESTION**
  To identify option questions, the classifier identifies this type of questions if these contain the conjunction "ou". For example, consider the question, "Gostas de carne ou peixe?". As the question contains the word "ou", this is classified as OR_QUESTION.

- **Etiqueta WH_QUESTION**
  The WH_QUESTION questions have rules based on regular expressions that indicate whether the questions begin with an adverb or an interrogative pronoun.

- **Etiqueta LIST_QUESTION**
  To identify this type of questions was developed a list of keywords that indicate/appear normally in this type of questions (for example: "Indica", "Indique", "Diz", "Diga", among others).

- In order to identify the dialog lines of type RHETORICAL_QUESTION, OPEN_ENDED_QUESTION, SOCIAL_OBLIGATIONS, ACKNOWLEDGEMENTS and NO_UNDERSTANDING was created a list (based on regular expressions) with expressions/sentences are marked according to the type of interaction in question.

### 6.2 Machine Learning Classifier

#### 6.2.1 Corpora

For the development of classifiers based on machine learning were created several corpora: corpus of personal/impersonal questions, corpus of factoids questions and corpus of questions/non-questions.

The corpus of personal and impersonal questions was built based on the corpus of personal questions used in Just.Chat [14]. This corpus is in English, and so it was necessary to make the translation into portuguese. The impersonal questions were taken from the corpus created by Li & Roth [11], who had already been translated into Portuguese [13]. The corpus contains 3710 questions, wherein 1756 are personal questions and 1954 are impersonal questions. This corpus was divided into training corpus (3340 questions) and test corpus (370 questions).

The corpus with factoids questions is a corpus with 6000 questions translated into Portuguese [13] (with their semantic categories) from the corpus in English created by Li & Roth [11].

For questions and non-questions, was developed a corpus with 4840 phrases, wherein 2420 are questions and 2420 are not questions.

#### 6.2.2 The classification process

The classifier used in this work is the classifier implemented in Just.Ask, which is based on a SVM, used one-versus-all version. The features used were as follows:

- **n-grams** - sequence of n consecutive words from a dialog line. We use the following n-grams: Unigram (U), Bigrams (B) e Trigrams (T).

- **binary n-grams** – it represents the presence (or absence) of n-grams in a dialog line through a binary value (0 or 1). The classifier uses binary-unigrams (BU), binary-bigrams (BB) e binary-trigrams (BT).

- **Length** (L) - indicates the number of tokens in a dialog line.

- **Word Shape** (WS) - indicates the number of tokens of a dialog line that are lowercase, uppercase, it have an uppercase first letter, those which are formed by digits, or, finally, that do not belong to any of the previous cases.

#### 6.2.3 The dialog line to its classification

It was developed two strategies for the implementation of the Classifier module. The first is based on a flowchart that consists of rule-based and based on machine learning classifiers. The flowchart in Figure 2 indicates the path taken to classify a dialog line, as well as the module that will be called, depending on the classification assigned to the dialog line.

The second strategy is based only on machine learning, base on features obtained with the previously proposed rules (section 6.1) and based on classifiers developed (section 6.2).

[5] http://www.cis.uni-muenchen.de/∼schmid/tools/TreeTagger/

### 6.3 Talkpedia Extentions

Talkpedia extensions were based on extensions carried out within the framework of the discipline of Natural Language of the 1st master year of the 1st semester, taught at Instituto Superior Técnico, in Taguspark campus, in the academic year 2013/2014. Students were asked to perform two mini-projects where they should apply the concepts learned in the discipline throughout the semester. The importance of these two mini-projects, in this thesis, due to the fact that both releted on the Talkpedia, in order to improve the performance of this and consist in the following:

- **First Mini-Project** In the first mini-projeto was asked the students to enrich the generic set of templates used by the Talkpedia and make them better suited to different types of interactions. After the analysis of some jobs, the best ideas we pulled are as follows: students have identified several types of interactions and classified them, and created a list of insults.

- **Second Mini-Project** In the second mini-projeto was asked the students to focus on processing and choice of responses. Again was made an analysis of some work carried out and on the processing step answers the main ideas that we got were:

  - Removal of footnote type information and acronyms
  - Normalization of whitespace.
  - Removal of "forgotten" characters as ', ,' or '–' and strange characters.
  - Deletion of phrases taken from Wikipedia that begin with *Nota:*.

  Finally, in step the choice of answers, ideas set aside for this phase were as follows:

  - Implement similarity measures (such as *Jaccard*, *Dice* and *EditDistance*) to search the various links obtained by Wikipedia.
  - To find the best phrase of a page for the answer to a given question, one of the solutions is to traverse all paragraphs of the page chosen and record that line where the search terms appear together, or those that appear more often in a particular phrase.

After the ideas taken from the extensions made by students, was developed a new extension of Talkpedia with all the features mentioned previously.

# 7 Evaluation

## 7.1 Evaluation of classifiers

To evaluate the performance of the classification used measures[6] *Precision*, *Recall*, *Accuracy* and $F_1 - Measure$, to calculate as follows (where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative):

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{3}$$

$$F_1 - Measure = 2.\frac{Precision.Recall}{Precision + Recall} \tag{4}$$

The evaluation method used was the following:

- In the evaluation of rule-based classification (section 7.2), the classifier based on a flowchart (section 7.4) and Mega-Classifier (section 7.5) was used the test corpora (6.1.1 and 6.2.1) using the measures referred to above.

- On evaluation of machine learning classifiers (section 7.3) was used the k-fold method cross validation, with a k equal to 10, and it was used the training corpora. For each iteration is calculated the measures mentioned above and at the end is calculated the average.

The corpora used were the previously described, in sections 6.1.1 and 6.2.1.

## 7.2 Evaluation of Rule-Based Classifiers

The evaluation of the Yes/No questions obtained a precision of 72,4%, a recall of 93,3% and a F-measure of 81,5%. In relation to the accuracy, the classifier was able to correctly classify 78,9% of questions. By analyzing in detail the results, the classifier was wrong in the questions of the list type, i.e. which begin, for example, with "Indique", classifying these as being YN_QUESTION. This happens because this type of questions starts with a verb (which is one of the patterns).

[6] http://en.wikipedia.org/wiki/Precision_and_recall

## 7.3 Evaluation of Machine Learning Classifiers

For the first strategy, were created four classifiers (isolated) based on automatic learning to identify the dialog lines type questions/non-question, yes/no questions, personal/impersonal questions and factoids questions (according to the classification of Li & Roth). To choose the best features for use in classifiers, these have been evaluated with the different features isolated and combined. For Yes/No questions, the best result was obtained by combining features unigrams and YN-Question (this feature was created by us, it is based on handwritten rules), with an accuracy of 96,6%, a precision of 97,6% and a recall of 97,6%.

The best result, for the identification of questions/non-questions, was obtained by the combination of features binary unigram, binary bigram and binary trigram, with an accuracy of 99,2%, precision and recall of 99,8% and F-measure of 99,3%.

For personal/impersonal questions the features unigram and word shape combined obtained the best result, with an accuracy of 95,1%, 94,7% of precision, a recall of 95,3% and F-measure of 95,0%.

The factoids questions, the best results for the fine classification for Li & Roth resulted from the combination of the features binary unigram and word shape, which obtained 66,8% of accuracy. For the coarse classification, the best combination were with the features binary unigram and word shape, with a accuracy of 78,9%.

## 7.4 TalKit Evaluation Based on a Flowchart

In the evaluation of classifier based on a flowchar, we used a test corpus consisting of 988 questions/phrases.

In General, the classifier had an accuracy of 87,3%, i.e. managed to rank well 863 questions/phrases; obtained an average precision of 83,4% and average recall of 83,7% (table 1).

| Tags | Precision | Recall |
|---|---|---|
| IMPERSONAL | 95,05% | 96,48% |
| PERSONAL | 73,55% | 82,41% |
| YN_QUESTION | 93,59% | 66,97% |
| DECLARATIVE_SENTENCE | 81,22% | 94,76% |
| LIST_QUESTION | 90,20% | **100,00%** |
| NO_UNDERSTANDING | 75,00% | 75,00% |
| RHETORICAL_QUESTION | 52,94% | **100,00%** |
| SOCIAL_OBLIGATIONS | 81,25% | 92,86% |
| OR_QUESTION | 91,67% | 73,33% |
| OPEN_ENDED_QUESTION | 87,50% | **100,00%** |
| ACKNOWLEDGEMENTS | **95,83%** | 38,33% |
| Mean | 83,44% | 83,65% |

Table 1: Results of precision and recall for each tag.

## 7.5 TalKit Evaluation Based on a Mega-Classifier

The `Mega Classifier` is a classifier based on machine learning, in which the features are composed by classifiers developed, referred to earlier. The training corpora used consists of 11981 questions/phrases taken from training corpora used in evaluations of previous classifiers. The test corpus is the same it was used in the previous section.

In General, the `Mega Classifier` managed to classify correctly 798 questions/phrases, i.e. 80,8% (less 7% than the classifier based on a flowchart). Obtained an average precision of 60,1% (less 23% compared to the previous classifier) and an average recall of 57.1% (less 26,5% compared to the previous classifier). One of the hypotheses for this classifier get worse results than the last, it may be the training corpus be unbalanced, i.e. there is more questions/phrases some of the types than others.

| Tags | Precision | Recall |
|---|---|---|
| IMPERSONAL | 95,05% | 95,76% |
| PERSONAL | 91,74% | 50,45% |
| YN_QUESTION | 0,00% | 0,00% |
| DECLARATIVE_SENTENCE | 84,08% | 91,56% |
| LIST_QUESTION | 90,20% | 95,83% |
| NO_UNDERSTANDING | 58,33% | 70,00% |
| RHETORICAL_QUESTION | 0,00% | 0,00% |
| SOCIAL_OBLIGATION | 87,50% | **100,00%** |
| OR_QUESTION | 58,33% | 87,50% |
| OPEN_ENDED_QUESTION | 0,00% | 0,00% |
| ACKNOWLEDGEMENTS | **95,83%** | 37,10% |
| Mean | 60,10% | 57,11% |

Table 2: Results of precision and recall for each tag.

## 7.6 TalKit Evaluation with Users

### 7.6.1 Experimental Setup

It was made a questionnaire, in order to realize from the developed classifiers we would direct a question to get a better answer in relation to use just the Say Something Smart or the Talkpedia.

The questionnaire was performed at 20 users and consists of 40 questions.

For each questions/phrases were presented between 2 to 5 answers (when there was equal answers was presented just one of them). The answers are given by the Say Something Smart, the Talkpedia in the original version, by the Talkpedia with the extensions implemented, by TalKit with the classifier based on a flowchart and the TalKit with `Mega Classificador`.

Users for each answer had to indicate whether this was plausible or not, then, had to order the answers in order of preference (from best to worst).

### 7.6.2 Results and Discussion

The obtained results are in tables 3 and 4, referring to answers plausible or not and for the best and worst answers.

From the table 3, the TalKit with classifier based on a flowchart obtained more plausible answers with an average of 14,5 answers and least plausible answers with an average of 5.5. The Original Talkpedia got less plausible answers with an average of 9,8 answers and more answers not plausible with 10,2.

The table 4 reinforces the conclusions obtained previously, where the worst responses were given by the Original Talkpedia (with 25 responses) and the best were given by TalKit with the classifier based on a flowchart (with 29 replies).

With the data taken from the questionnaire it is also possible to draw conclusions about the modifications made to the Talkpedia. According to the table 3, the Talkpedia with extensions obtained more plausible responses (with an average of 11,3 answers) than the Original Talkpedia (with an average of 9,8 answers). And according to table 4, the Talkpedia with extensions obtained best answers (with 16 best answers and only 14 worse answers) than the original Talkpedia (with 13 best answers and 25 worst answers). Therefore, we can conclude that, effectively, the extensions made to the Talkpedia improved the answers.

# 8 Conclusions

Dialog systems are often taken to the limit by users to test the capacity of the same. For this reason, it is necessary to develop platforms able to answer any question. This thesis was developed the TalKit, a platform, which on the basis of a classifier, allows it to choose between different modules that provide a answer to a user using the classifier taxonomy proposed in this thesis.

According to the proposal taxonomy, two strategies were developed for the implementation of the TalKit classifier. The first is based on a flowchart where are combined handwritten rules and machine learning. The second strategy is based only on machine learning (the `Mega Classificador`), and uses as features previously developed rules, as well as the remaining classifiers used in the flowchart.

It had been tested two classifiers, a flowchart-based and the other in machine learning. In the evaluation of these two strategies, it was concluded that the first is better than the second, as it managed to classify 87% of interactions with success (the second strategy managed to classify only 80% of interactions with success).

An evaluation is made with users, it was concluded that using the TalKit based on aflowchart gets more plausible responses (14 answers plausible) than using only the Talkpedia (original with 9,8 plausible answers and extension with 11 plausible answers) or theSay Something Smart (with plausible answers 10,6).

# 9 Future Work

There are several tasks you can perform in the future with the aim of improving the TalKit. Here are the suggestions for future work:

- Integrate other systems (e.g. Just.Ask).
- Enrich the corpus of yes/no questions, rhetorical questions, open-answer questions, questions of list type, expression phrases greetings and goodbyes, and phrases that indicate confirmation.
- Keep the history of the conversation and be able to choose a best answer according to the context.
- For questions/phrases of type NO_UNDERSTANDING the system be able to rephrase the answer given earlier.
- Enrich the rules created for the classifier based on rules.

# References


[1] J. Allen and M. Core. Draft of DAMSL: Dialog act markup in several layers. Unpublished manuscript, 1997.

[2] D. L. S. Ameixa. Say something smart. Master's thesis, Instituto Superior Técnico, Nov. 2013.

[3] L. Becker, W. H. Ward, S. van Vuuren, and M. Palmer. Discuss: A dialogue move taxonomy layered over semantic representations. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, pages 310–314, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[4] H. Bunt. Context and dialogue control. *Think*, 3:19–31, 1994.

[5] H. Bunt. The dit++ taxonomy for functional dialogue markup. In D. Heylen, C. Pelachaud, R. Catizone, and D. Traum, editors, *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24, 2009.

| | **Plausible** | **Non plausible** | **Total** |
|---|---|---|---|
| Classificador Fluxograma | **14,475** | 5,525 | 20 |
| Mega Classificador | 10,950 | 9,050 | 20 |
| Say Something Smart | 10,625 | 9,375 | 20 |
| Original Talkpedia | 9,825 | **10,175** | 20 |
| Talkpedia with extensions | 11,300 | 8,700 | 20 |
| Mean | 11,435 | 8,565 | 20 |

Table 3: The average number of plausible answers and not plausible for classifier and system.

| | Flowchart Classifier | Mega Classificador | Say Something Smart | Original Talkpedia | Talkpedia with extensions |
|---|---|---|---|---|---|
| Melhor | **29** | 18 | 16 | 13 | 16 |
| Pior | 3 | 14 | 14 | **25** | 14 |

Table 4: Number of best and worst answers by classifiers and systems.


[6] M. G. Core and J. F. Allen. Coding Dialogs with the DAMSL Annotation Scheme. In *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*, Boston, MA, November 1997.

[7] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg. Meeting recorder project: Dialog act labeling guide. Technical report, Technical Report TR-04-002, ICSI, 2004.

[8] P. Fialho, L. Coheur, S. dos Santos Lopes Curto, P. M. A. Cláúdio, A. Costa, A. Abad, H. Meinedo, and I. Trancoso. MEET EDGAR, A TUTORING AGENT AT MONSERRATE. In *ACL*, Proceedings of the 51st Annual Meeting of the Association f, Aug. 2013.

[9] ISO. 24617-2 (2010) language resource management, semantic annotation framework (semaf), part 2: Dialogue acts. Technical Report ISO 24617-2, 2010.

[10] D. Jurafsky, E. Shriberg, and D. Biasca. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. Technical Report Draft 13, University of Colorado, Institute of Cognitive Science, 1997.

[11] X. Li and D. Roth. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[12] P. Mota. LUP: A language understanding platform. Master's thesis, Instituto Superior Técnico, jul 2012.

[13] Ângela Costa, T. Luís, J. Ribeiro, A. C. Mendes, and L. Coheur. An english-portuguese parallel corpus of questions: translation guidelines and application in smt. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[14] M. J. d. M. M. P. Pereira. Just.chat - dos sistemas de pergunta/resposta para os chatbots. Master's thesis, Instituto Superior Técnico, May 2013.

[15] R. M. P. Pires. Query classification and expansion in just.ask question answering system. Master's thesis, IST/Universidade Técnica de Lisboa, June 2012.

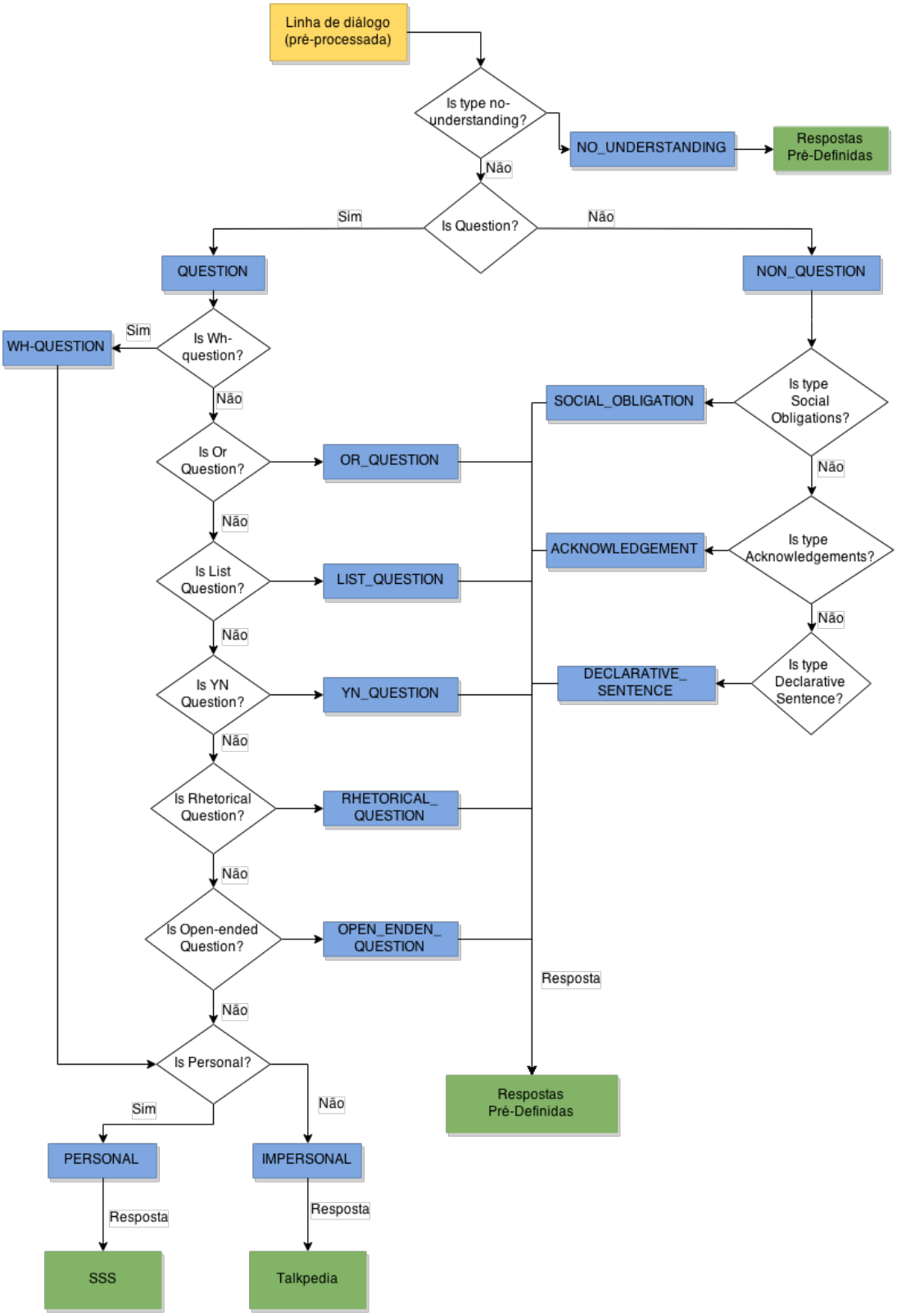


Figure 2: Flowchart concerning the choice of classification and response.