Distinguishing Different Classes of Utterances – the UC-PT Corpus

Mariana Gaspar Fernandes
INESC-ID, Lisboa, Portugal
Instituto Superior Técnico, Universidade de Lisboa, Portugal
mariana.gaspar.fernandes@tecnico.ulisboa.pt

Cátia Dias
INESC-ID, Lisboa, Portugal
Instituto Superior Técnico, Universidade de Lisboa, Portugal
catiadias089@gmail.com

Luísa Coheur
INESC-ID, Lisboa, Portugal
Instituto Superior Técnico, Universidade de Lisboa, Portugal
luisa.coheur@tecnico.ulisboa.pt

Abstract

Conversational bots are being used in many scenarios and we can find them playing museum guides or providing customer support, for instance. These bots base their answers in specific information related with their domain of expertise, but there is general information, presented in each user request that, when properly identified, could also be useful for the agent to decide what to answer. As an example, if the user is asking a question or uttering a statement, the bot's action in its search for a response will probably differ. In this paper we present three corpora for the Portuguese language – the UC-PT corpus – that can be used to help conversational bots to distinguish: a) questions from non questions, b) yes-no-questions from other types of questions; and c) personal from non-personal questions. With this information, the agent can decide, for instance, not to answer, redirect the question to a persona chatbot or decide to answer it with a simple "yes", "no" or "maybe". In addition, we benchmark the classification process in these corpora. This corpora will be made publicly available.

2012 ACM Subject Classification Information systems → Question answering; Computing methodologies → Language resources; Social and professional topics → Computer and information systems training; Applied computing → Annotation; Computing methodologies → Supervised learning; Information systems → Information extraction

Keywords and phrases Corpora, Questions, Conversational Agents, Portuguese Language

Digital Object Identifier 10.4230/OASIcs.SLATE.2019.14

Funding This work was supported by national funds through Fundação para a Ciência e Tecnologia (FCT) with reference UID/CEC/50021/ 2019 and by FCT's INCoDe 2030 initiative, in the scope of the demonstration project AIA, "Apoio Inteligente a Empreendedores (chatbots)".

1 Introduction

Conversational bots (often called chatbots) have been accompanying the recent advances in Artificial Intelligence (AI). We can find them in many different scenarios [11], such as in museums or providing customer support: Edgar Smith [6], the Monserrate's Palace butler illustrates the former; IKEA's Anna, the latter. Although some recent conversational agents are data-driven, and take advantage of the latest advances in Deep Learning (e.g. [12]), they also need large quantities of data to be trained. Thus, in most scenarios this approach is not

8th Symposium on Languages, Applications and Technologies (SLATE 2019).
Editors: Ricardo Rodrigues, Jan Janoušek, Luís Ferreira, Luísa Coheur, Fernando Batista, and Hugo Gonçalo Oliveira; Article No. 14; pp. 14:1–14:8

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

possible. Several platforms, as for instance RASA[1], contribute to a faster development of chatbots, but, still, much manual work is needed, as corpora needs to be provided, so that the bots can learn, for instance, to identify the user intentions or the entities mentioned in his/her utterances. If the data needed to train these systems is usually dependent of the application domain, there is general and useful information that can be extracted from the user requests, which can help the chatbot to further decide how to answer, namely to identify if the user request is a question or not, if the user is posing a personal question to the agent or not, or if the question can be answered with a yes/no/maybe or not. Considering that such resource could be useful for the Portuguese community developing conversational bots, we propose, in this work, the UC-PT corpus, which is constituted of the following corpora:

- the Question vs. Non-question corpus: a corpus with 5034 utterances labeled as "question" (e.g., "O que são minhocas de pesca?" – "What are fishing worms?"), and "non-question" (e.g., "Precisaremos de manter a lei e a ordem, se isto acontecer." – "We will need to maintain law and order, if this happens.");
- the Personal vs. Impersonal corpus: a corpus with 3698 utterances labeled as "personal" (e.g., "Quem é a pessoa mais calma que conheces?" – "Who is the calmest person that you know?"), and "impersonal" (e.g., "Que diferentes tipos de plástico existem?" – "How many different kinds of plastic are there?");
- the Yes/No questions corpus: a corpus with 360 utterances labeled as "yes/no" questions (e.g., "Gostas de cangurus, querida?" – "Do you like kangaroos, sweetheart?"), and "other" (e.g., "Quantos cardeais elegem o Papa?" – "How many cardinals elect the pope?").

After building and annotating these three corpora, we calculate the inter-annotator agreement with Cohen's kappa score, obtaining two near-perfect and one perfect agreement. In addition, we benchmark the classification process. By using simple features as n-grams, we get accuracies ranging from 97% to 100%.

This paper is structured as follows: in Section 2 the three aforementioned corpora are described, and, in Section 3, we explain what we did to classify the different corpora, and which solutions provided the best results. Finally, in section 4 we present our conclusions and point to some future work.

2 Building Corpora

The sentences in all the three corpora came from different sources, namely:

- from the translation into Portuguese [5] of the widely used Li & Roth corpus [7];
- from a manual Portuguese translation of parts of the corpora of a chatbot called JustChat [10];
- from the B-Subtle corpus, a corpus built from movies Subtitles, as described in [3].

In addition, some sentences were gathered from the web, created by the authors of this work or suggested by Técnico Students at Taguspark, in a Natural Language course.

In the following we describe each one of the three corpora.

2.1 Question vs. Non-Question corpus

In this section we give a brief description of the Question vs. Non-question corpus, namely, the different formulations of questions that were gathered, as well as some examples of the non-questions.

[1] `https://rasa.com`

2.1.1 Questions

Several types of questions were taken into consideration. Besides the usual *direct questions*, we also gave some room to *imperative sentences*[2] that constitute a request for information.

Regarding *direct questions*, the corpus contains several examples of the so called "Wh-questions", that is, questions that contain the keywords "quem" ("who"), "onde" ("where"), "porquê, porque" ("why"), "o quê, qual" ("what"), "o quê, qual" ("which"), "quando" ("when") and "como" ("how"). Examples of such questions are:

- "Quem é Alan Turing?" – "Who is Alan Turing?";
- "Qual é o nome abreviado do Mississippi?" – "What is the nickname for the state of Mississippi?";
- "O que é a viscosidade?" – "What is viscosity?";
- "Quando foi travada a batalha de Somme?" – "When was the battle of Somme fought?";
- "Por que motivo foi inventado o fecho de correr?" – "Why was the zipper invented?".

We also included in the corpus wh-questions that present a possibility, like an imagined scenario, and then inquire something with that scenario in mind (e.g., "Se o mundo inteiro estivesse ouvindo, que dirias?" – "If the whole world was listening, what would you say?"). In addition, we also addressed questions whose answer can be a simple "Sim" ("Yes") or "Não" ("No") (e.g., "Gostas de ler?" – "Do you like to read?" or "Andas na escola?" – "Do you go to school?"), including questions that are only one word (e.g., "Jantar?" – "Dinner?"). Moreover, questions that contain two possibilities of answer separated by the connector "or" (choice questions), were also added to the corpus (e.g., "Do que gostas mais: factos ou ficção?" – "What do you like more: facts or fiction?").

In what concerns *imperative sentences* that constitute a request for information, or ask for a description or definition of something, several cases were included in the corpus. Some examples are:

- "Mencione um cetáceo." – "Mention a cetacean.";
- "Diga o nome da organização que é presidida por um Conselho de Segurança." – "Say the name of the organization that is presided by a security counsel.";
- "Descreva a aparência do músico Finlandês Salonen." – "Describe the Finnish music personality Salonen's appearance.";
- "Defina cosmologia." – "Define cosmology.".

Finally, we opted to add some cases where several questions are formulated in the same entry. The reason for this is that sometimes people ask several questions related to each other in a row (e.g., "Se tivesses de escolher, qual animal de uma quinta gostarias ser? Porquê? Podes fazer o som?" – "If you had to pick, which farm animal would you like to be? Why? Can you do its sound?").

In summary, the utterances labeled as questions encompass: choice questions, wh-questions, yes/no questions and imperative sentences.

2.1.2 Non-questions

The non-questions part of the corpus is constituted of sentences such as:

- "A ideia é os dez formarem um círculo de protecção em torno do possuído." – "The idea is that the ten form a circle around the possessed.";

[2] Sentences that are an order, an instruction or a request to do something [1].

- "O David precisa de ir ao lançamento de um filme." – "David needs to go to a movie launch.";
- "Deixou a faculdade de direito, não tem emprego." – "Left law school, has no job.";
- "Não é motivo para renegar a família." – "It is no motive to renegate the family.".

2.1.3 Some Statistics

We randomly split the corpus in two, one part for training and one part for testing. The training corpus contains 4526 entries, from which 2280 are labeled as "question" and the remaining 2246 as "non-question". The testing corpus contains 508 entries from which 264 are "non-questions" and 244 are "questions". Extra details can be found on Table 1[3].

Table 1 Statistics about the Question vs. Non-question corpus.

	Training Set	**Testing Set**	**Whole Corpus**
Number of Tokens	42614	4552	47166
Number of Unique words	7741	1509	8253
Average Word Length	4.26	4.31	4.26
Number of Characters	162610	17464	180074
Number of StopWords	13651	1442	15093
Number of Words	36812	3901	40713

2.2 Personal vs. Impersonal Questions

In this section we explain what can be found in the Personal vs. Impersonal corpus.

2.2.1 Personal Questions

In European Portuguese, the way personal questions are formulated depend on who we are talking to. Hierarchy and age difference, among others, will lead to more formal/informal conversations. When two people engage in an informal conversation, the second person of the singular is usually used; otherwise the third person of the singular is employed. For instance, if we ask a friend if he likes to read, we would ask "Gostas de ler?" or "Tu gostas de ler?", but if we asked a person we do not know or has one of the aforementioned differences, we would ask "Gosta de ler?" or "Você gosta de ler?" (being the latter in a more Brazilian Portuguese style). In the English language all these questions translate to "Do you like to read?". In the corpus for personal and impersonal questions we took these cases into account. Examples are (the first one is an example of formal speech, and the second of informal speech):

- "Diga algo que fez em criança que os seus pais não sabem." – "Say something that you did as a child that your parents do not know of.";
- "Diz 1 coisa que desejavas mudar em ti." – "Say 1 thing that you wish to change in yourself.".

Other examples of personal questions that can be found in the corpora are related with: a) situations in which the user presents a scenario and then asks what the other person would do considering it (e.g., "Se tivesses que comer um guaxinim ... como irias cozinhá-lo?" – "If

[3] In this and in the remaining corpora, the number of tokens and the number of characters take into consideration punctuation.

you had to eat a raccoon... how would you cook it?"); b) personal preferences (this can be regarding to movies, food, among other personal tastes) (e.g., "Qual é o teu filme favorito?" – "What is your favourite movie?"); c) family, friends, romantic relationships, among others (e.g., "O que me podes dizer sobre um dos teus avós?" – "What can you tell me about one of your grandparents?"); d) feelings, opinions, beliefs and visions in life: (e.g., "Achas que é correto namoriscar se tens namorado/namorada?" – "Do you think it is ok to flirt if you have a boyfriend/girlfriend"); e) past and/or a person's experience (e.g., "Indica 1 coisa que te faz falta das férias quando eras criança." – "State 1 thing that you miss of the vacations you had when you were a child."); f) what a person wears and his/her appearance, habits, skills, personal info/data, personal options, facts about personal life, etc. (e.g., "És bom a escrever na tua língua materna?" – "Are you any good at writing in your mother tongue?"). In conclusion, personal questions are questions about the interlocutor's personal matters, such as his opinions, feelings, memories, home city, friends, among others. If the questions are about the personal life of a person that is not an acquaintance of the interlocutor and if that question is not asking for an opinion, then it is not personal.

2.2.2 Impersonal Questions

As for the impersonal questions, they are mostly factoid questions extracted from the aforementioned translation of Li & Roth corpus for Portuguese. Some examples include:

- "O que faz com que um tornado gire?" – "What makes a tornado turn?";
- "Quais são os dois países cuja costa faz fronteira com a Baía de Biscaia?" – "What two countries' coastlines border the Bay of Biscay?";
- "Que actor casou com a irmã de John F. Kennedy?" – "What actor married John F. Kennedy's sister?".

2.2.3 Some Statistics

The personal/impersonal training corpus has 3329 queries, from which 1746 are labelled as "impersonal" and the other 1583 are labelled as "personal". The testing corpus has 369 entries from which 205 are tagged as "impersonal" and the other 164 are tagged as "personal". More detailed statistics about this corpus can be found on Table 2.

Table 2 Statistics about the personal and impersonal corpus.

	Training Set	**Testing Set**	**Training + Testing Set**
Number of Tokens	33407	3733	37140
Number of Unique words	5714	1173	6099
Average Word Length	4.38	4.29	4.37
Number of Characters	132257	14413	146670
Number of StopWords	10197	1117	11314
Number of Words	29272	3248	32520

2.3 Yes/No Questions vs. Other

In this section we explain what are Yes/No questions and we provide some examples of the questions of this kind that can be found in this corpus. We also present some examples of the questions that cannot be answered with a simple "yes", "no" or "maybe".

2.3.1 Yes/No Questions

Examples of Yes/No questions are:

- "Lês muito?" – "Do you read a lot?";
- "Gostas de dançar?" – "Do you like to dance?";
- "Tens dinheiro?" – "Do you have money?";
- "Ontem choveu?" – "Did it rain yesterday?".

Notice that, in the set of Yes/No questions, one can find questions constituted of one single word (e.g., "Pizza?").

2.3.2 Other

As to the questions labeled as other, they are like the ones presented in Section 2.1, excluding the Yes/No ones. Under the label "other" we can find questions such as "Wh-questions", imperative sentences, among others. Here are some examples (extracted from the corpus):

- "Indique um pesticida." – "State a pesticide.";
- "Em que cidade se encontra a Basílica de São Pedro?" - "In what city is Saint Peter's basilica located?";
- "És de que clube?" – "Of what club are you?";
- "Quanto custou o Túnel da Mancha?" – "How much did the channel tunnel cost?".

2.3.3 Some Statistics

The training corpus has 320 entries, from which 157 are labeled as "yes/no-question" and the other 163 as "other". As for the testing corpus it contains 40 entries from which 19 are labelled as "other" and the other 21 are labelled as "yes/no-question". More detailed information can be found in Table 3.

Table 3 Statistics about the Yes/No Question and Other corpus.

	Training Set	**Testing Set**	**Training + Testing Set**
Number of Tokens	2058	261	2319
Number of Unique words	723	147	787
Average Word Length	4.59	4.69	4.60
Number of Characters	8199	1057	9256
Number of StopWords	530	65	595
Number of Words	1711	216	1927

2.4 Inter-annotator Agreement

A random sample of 100 queries was selected from each of the above corpus, rendering for each corpus 50 queries for each label. This sample was given to three different annotators (one for each corpus) which in turn gave their annotation for each query. Upon doing this, the results were compared with the original labelling, made by a single annotator, using the Cohen's kappa coefficient metric (using the implementation provided by scikit-learn [9]). The results obtained can be found in Table 4, and show that, for the Question vs. Non-question corpus, there is a perfect agreement between the two annotators. As for the other two corpus there is a near-perfect agreement.

Table 4 Inter annotator agreement results.

Corpus	Cohen Kappa Score
Question and Non-Question	1.00
Personal and Impersonal	0.88
Yes/No and Other	0.98

Some examples of sentences in which the annotators did not agree in the Personal vs. Impersonal corpus are:

- **impersonal**: "Porque estamos na Terra?" – "Why are we on Earth?";
- **personal**: "Quando saem os objectos de Halloween nas lojas no teu país?" – "When do the Halloween objects come out in your country's stores?".

As these questions could be answered with both opinions and facts, it is understandable that that ambiguity causes a non-agreement between the two annotators.

The only sentence in which the annotators did not agree in the Yes/No question vs. Other corpus was:

- **yes/no-question**: "Do Stephen King? Um filme de terror?" – "From Stephen King? An horror movie?".

This question could be both answered with a simple yes or no, and with a movie. Which explains why the annotators did not agree on the label.

3 The Classification Process

We conducted the classification process by creating models with Naive Bayes (NB) and Support Vector Machines (SVM). We used NB due to its simplicity and SVM for its proven effectiveness in the task of text classification, as discussed in [2]. We have experimented with all the three Vectorizers (CountVectorizer, HashingVectorizer and TfidfVectorizer), with all implementations of NB (Complement, Bernoulli, Multinomial and Gaussian), and with SVM where we have tried 4 distinct kernels (linear, rbf, sigmoid and poly). As features we used Unigrams, Bigrams, Trigrams and combinations of them. We have also tested with a custom tokenizer (the TweetTokenizer available in the Natural Language Toolkit [4]). The evaluation metric that we used was accuracy. The best results ranged from 98.10% to 100% (accuracy) in the three corpus for the aforementioned train/test partitions. TweetTokenizer led to the best results in all the corpora. 100% accuracy was obtained for both the Question/Non-Question and the Yes/No question corpus, with SVM + linear kernel, CountVectorizer and Unigrams as features. As for the Personal vs. Impersonal corpus, the obtained accuracy was of 98.10%, with SVM + linear kernel, TfidfVectorizer, and Unigrams + Bigrams or Unigrams + Bigrams + Trigrams as features. For the previous experiment, NB led to similar results (Complement).

Additionally, we performed a 10-fold cross-validation on each corpus (training and testing corpus together) using the classification pipeline that presented the best results for each corpus in the train/test partition classification. With these pipelines, we obtained a range of accuracies between 97% and 100%.

Results show that it is not complicated to discriminate between the proposed different types of utterances.

Although we do not want to impose an order in the usage of these corpora, an obvious scenario is: first, the model trained in the Question vs. Non-Question corpus is used to check whether a query is or is not a question. If it is a question it can be further classified as a yes/no question vs. other and, in addition, as personal or impersonal.

4 Conclusions and Future Work

We presented the UC-PT corpus (which will be made available upon request), which is built on three different corpora: one that is constituted of questions and non-questions, one that has personal and impersonal questions, and the last one that comprises questions that can and cannot be answered with a yes/no/maybe. We had very high inter-annotator agreements with the three corpora. We also tested several classifiers and obtained accuracies higher than 97%. As future work we would like to use the created models to improve a conversational bot as well as create new corpora or rules to get a classification of a sub-type of questions or non-questions (wh-questions, declarative sentences, or-questions, and so on and so forth). Additionally, the Question vs. Non-Question corpus can be enriched with indirect questions such as "Pergunto-me qual é a capital da Finlândia." – "I wonder what is the capital of Finland." (could also be added to the Personal vs. Impersonal corpus) which indirectly ask for an answer about something. Finally, we will focus on answers' classification, and, in particular, we will explore how to take advantage of relations between answers [8].

References

1 Bas Aarts. *Oxford Modern English Grammar*. Oxford University Press, 2011.

2 Charu C. Aggarwal and ChengXiang Zhai. *A Survey of Text Classification Algorithms*, pages 163–222. Springer US, Boston, MA, 2012.

3 David Ameixa. Say Something Smart - ensinando um chatbot a responder com base em legendas de filmes. Master's thesis, Instituto Superior Técnico, Lisboa, Portugal, 2015.

4 Edward Loper Bird, Steven and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

5 Ângela Costa, Tiago Luís, Joana Ribeiro, Ana Cristina Mendes, and Luísa Coheur. An English-Portuguese parallel corpus of questions: translation guidelines and application in SMT. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association (ELRA), 2012.

6 Pedro Fialho, Luísa Coheur, Sérgio dos Santos Lopes Curto, Pedro Miguel Abrunhosa Cládio, Ângela Costa, Alberto Abad, Hugo Meinedo, and Isabel Trancoso. MEET EDGAR, A TUTORING AGENT AT MONSERRATE. In *ACL*, Proceedings of the 51st Annual Meeting of the Association f, August 2013.

7 Xin Li and Dan Roth. Learning Question Classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi:10.3115/1072228.1072378.

8 Ana Cristina Mendes and Luísa Coheur. An Approach to Answer Selection in Question-Answering Based on Semantic Relations. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 1852–1857, 2011. doi:10.5591/978-1-57735-516-8/IJCAI11-310.

9 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

10 Maria Pereira. Just.chat - dos sistemas de pergunta/resposta para os chatbots. Master's thesis, Instituto Superior Técnico, Lisboa, Portugal, 2015.

11 Maria João Pereira, Luísa Coheur, Pedro Fialho, and Ricardo Ribeiro. Chatbots' Greetings to Human-Computer Communication. *CoRR*, abs/1609.06479, 2016. arXiv:1609.06479.

12 Yiping Song, Rui Yan, Yansong Feng, Yaoyuan Zhang, Dongyan Zhao, and Ming Zhang. Towards a Neural Conversation Model With Diversity Net Using Determinantal Point Processes. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, 2018.