Lexical Analysis and Content Extraction from Customer-Agent Interactions

Sergiu Nisioi **Anca Bucur*** **Liviu P. Dinu**
Human Language Technologies Research Center,
*Center of Excellence in Image Studies,
University of Bucharest
{sergiu.nisioi,anca.m.bucur}@gmail.com, ldinu@fmi.unibuc.ro

Abstract

In this paper, we provide a lexical comparative analysis of the vocabulary used by customers and agents in an Enterprise Resource Planning (ERP) environment and a potential solution to clean the data and extract relevant content for NLP. As a result, we demonstrate that the actual vocabulary for the language that prevails in the ERP conversations is highly divergent from the standardized dictionary and further different from general language usage as extracted from the Common Crawl corpus. Moreover, in specific business communication circumstances, where it is expected to observe a high usage of standardized language, code switching and non-standard expression are predominant, emphasizing once more the discrepancy between the day-to-day use of language and the standardized one.

1 Introduction

It is often the case for companies that make use of a customer relationship management software, to collect large amounts of noisy data from the interactions of their customers with human agents. The customer-agent communication can have a wide range of channels from speech, live chat, email or some other application-level protocol that is wrapped over SMTP. If such data is stored in a structured manner, companies can use it to optimize procedures, retrieve information quickly, and decrease redundancy which overall can prove beneficial for their customers and maybe, more important, for the well-being of their employees working as agents, who can use technology to ease their day-to-day job. In our paper, we work with email exchanges that have been previously stored as raw text or html dumps into a database and attempt bring up some possible issues in dealing with this kind of data lexically, from an NLP perspective, but also to forward a solution for cleaning and extracting useful content from raw text. Given the large amounts of unstructured data that is being collected as email exchanges, we believe that our proposed method can be a viable solution for content extraction and cleanup as a preprocessing step for indexing and search, near-duplicate detection, accurate classification by categories, user intent extraction or automatic reply generation.

We carry our analysis for Romanian (ISO 639-1 ro) - a Romance language spoken by almost 24 million people, but with a relatively limited number of NLP resources. The purpose of our approach is twofold - to provide a comparative analysis between how words are used in question-answer interactions between customers and call center agents (at the corpus level) and language as it is standardized in an official dictionary, and to provide a possible solution to extract meaningful content that can be used in natural language processing pipelines. Last, but not least, our hope is to increase the amount of digital resources available for Romanian by releasing parts of our data.

2 Data

While acknowledging the limits of a dictionary, we consider it as a model of standardized words, and for this we make use of every morphological form defined in the Romanian Explicative Dictionary DEX[1] - an electronic resource containing both user generated content and words normed by the Romanian Academy. We extract from the database a total of over 1.3 million words including all the morphological inflected forms. It is important to note here, the user generated content is being curated by volunteers and that not every word appearing in the dictionary goes through an

[1] https://dexonline.ro

Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text, pages 132–136
Brussels, Belgium, Nov 1, 2018. ©2018 Association for Computational Linguistics

official normative process for the language. In consequence, this resource may contain various region-specific word forms, low frequency or old terms and other technical neologisms.

While a dictionary can provide the list of words, it certainly lacks context and the way language is used in a large written corpus. One of the largest corpora of Romanian texts consists of news articles extracted from Common Crawl[2], it consists of texts on various topics and genres, and recently it has been considered (Bojar et al., 2016) a reliable resource for training a generic language model for modern standard Romanian, as part of the News task, Workshop of Machine Translation 2016. This corpus contains 54 million words, it covers general content not related to a specific topic, and since it has been scraped from public websites, it is reasonable to assume it contains standard official Romanian text, grammatically and lexically correct.

The question-answer corpus consists of interactions saved from an Enterprise Resource Planning (ERP) environment within a private Romanian company. All data has been anonymized before usage and personally identifiable information has been removed. The topics are highly business-specific, covering processes such as sales, human resources, inventory, marketing, and finance. The data consists of interactions in the form of tasks, requests, or questions (Q) and activities, responses, or answers (A).

One question may have multiple answers and the documents may contain email headers, footers, disclaimers or even automatic messages. To alleviate the effect of noise on our analysis, we have implemented heuristics to remove automatic messages, signatures, disclaimers and headers from emails.

	questions	**answers**
# tokens	7,297,400	11,370,417
# types	4,425,651	4,439,299
type / token ratio	0.6065	0.3904
total tokens	18,667,817	

Table 1: Question answering corpus size

The statistics regarding the size of the corpus are rendered in Table 1, we observe that the number of types (unique words) is quite similar for both questions and answers, however the total number of words used in the responses is a magnitude larger than the one corresponding to questions. Considering that type to token ratio is a reasonable indicator for lexical richness (Read, 2000), then customers use a rich vocabulary to describe their problems, with a considerable high probability for new words to appear in the received queries, while agents show a more standardized, smaller vocabulary to formulate their replies.

[2] http://commoncrawl.org

	Questions	**Answers**	**Common Crawl**
Vocabulary size	21,914	25,493	148,980
Dict diacr. overlap	41.75	40.65	42.22
Dict no diacr. overlap	55.51	52.96	60.87
Answers overlap	67.87	-	10.59
Answers diff English	4.83	-	7.28
Questions overlap	-	58.34	8.96
Questions diff English	-	20.95	8.04
C. Crawl overlap	60.9	61.86	-
C. Crawl diff English	7.13	13.17	-

Table 2: Comparison of overlapping dictionaries

3 Quantitative Lexical Analysis

We carry a comparison at the lexical level, in particular by looking at the size and variety of the vocabulary with respect to a standard Romanian dictionary. We extract word2vec embeddings [3] using CBOW with negative sampling (Mikolov et al., 2013; Řehůřek and Sojka, 2010) for three corpora: Common Crawl, the corpus of Questions, and the one containing Answers. The models are trained to prune out words with frequency smaller than 5, shrinking the vocabulary to ensure that the words included have good vectorial representations. From those vocabularies, we discard numbers, punctuation and other elements that are not contiguous sequences of characters.

We then proceed to use the vocabulary from the trained models and compare against the entire dictionary Romanian of inflected forms. For the later, we build two versions - one which contains diacritics and a second one which contains words both with and without diacritics.

For each vocabulary at hand we perform two simple measurements:

1. **overlap** - the percentage of overlap between one vocabulary and another

2. **diff English** - the percentage of differences between one vocabulary and another, that are part of an English WordNet synset (Fellbaum, 1998)

[3] The resources are released at https://github.com/senisioi/ro_resources

These basic measurements should give an indicator on how much of the vocabulary used in our ERP data is covered by the generic resources available for Romanian, and how important *domain adaptation* is for being able to correctly process the texts.

Table 2 contains the values for these measurements in a pair-wise fashion between each vocabulary - dictionary with and without diacritics, questions and answers vocabulary, and Common Crawl model vocabulary. We also compare the vocabularies extracted from our corpora with the dictionary having diacritics removed, as it is often the case to write informal Romanian with no diacritics. The second and third rows show an increase in overlapping percentage, regardless of the vocabulary, when the diacritics are ignored, which indicates that even official news articles contain non-standard words and or omissions of diacritics. It is, therefore, expected that a highly technical domain such as business-finance to have an even smaller overlap with the standard dictionary.

Somewhat surprising is the fact that a big majority of words from the Common Crawl vocabulary (approx. 39%) is not available in the full dictionary, and at a closer look we observe that 11.01% of words are also part of the English WordNet synsets (Fellbaum, 1998).

Furthermore, both the lexicons used in questions and answers present little overlap with Common Crawl, and in accordance with the lexical richness evidenced in Table 1, we observe that the vocabulary specific to answers overlaps better with the one for questions than vice-versa. In addition, over 20% of the words used in questions that do not appear in the answers are part of an English WordNet synset.

While the language of questions and answers is used in a business environment, one expecting it to be more formal and closer to the standard, the contrary appears to be true - to improve the speed of communication, people prefer to code switch between Romanian and English, not to use diacritics at all or to insert abbreviations and foreign words adapted to the Romanian morphology (e.g., *loga*, *loghez*, verb, used as in English *to log* or *to log in* most similar to dictionary verbs *to connect* and *to authenticate*).

A few examples of queries from the models are rendered in Table 4, showing that the domain-specific models learn good representations for abbreviations of specific terms (e.g., *exemplu (example) - ex, factura (invoice) - fact, fc*) being more robust to noise and free-form lexical variations in language.

	questions	**answers**
function words	17.22	16.47
pronouns	5.11	4.78
sentences	14.81	11.29
token length	4.91	5.27

Table 3: Average number of features / question or answer

At last, in Table 3, we count the average number of content independent features (function words, pronouns, number of sentences and average token length) that appear in both questions and answers. These features can provide information regarding the style of a text (Chung and Pennebaker, 2007), being extensively used in previous research for authorship attribution or literary style identification (Mosteller and Wallace, 1963; Koppel et al., 2009). Here we observe a stylistic difference between how questions and answers are formulated, questions being longer and more complex, which can also be a reason behind the smaller average length of the tokens, as Menzerath-Altmann law (Altmann and Schwibbe, 1989) states - the increase of a linguistic construct determines a decrease of its constituents.

4 Content Extraction

The lexical analysis in the previous section strongly suggests that our question-answering corpus contains a high vocabulary richness that is non-standard and divergent from the generic resource available for Romanian. Therefore, any type of text processing from classification, retrieval, or tagging is error prone and can provide misleading results. An important step, is therefore, to detect and extract the relevant content that best explains the customer's intent, which can be further used for classification or automatic reply generation.

Having very few resources at disposal, we proceeded to build our own dataset for intent extraction and annotated approximately 2000 requests, having in total 200,000 words. For each document at hand, we remove the sentences that did not contain relevant content and created aligned document-to-document pairs consisting of the full document to the left and the relevant content to the right. More exactly, the annotations are being

word	Q/A model	score	C. Crawl model	score
pentru (for)	pt	0.85	pt	0.61
	Pentr	0.74	nevoie (need)	0.49
	ptr	0.64	special	0.49
ex (for example)	exemplu (example)	0.77	676	0.78
	Ex	0.65	pixuletz (pen)	0.78
	Exemplu	0.65	dreaming	0.78
	adica (which means)	0.6	thd	0.78
banca (bank):	registru (register)	0.79	autoritatea (authority)	0.79
	numerar (cash)	0.78	lege (law)	0.78
	casa (cash desk)	0.73	nationala (national)	0.78
	plati (payments)	0.73	reforma (reform)	0.77
factura (invoice)	facture	0.84	lunara (monthly)	0.85
	comanda (order)	0.77	pompa (pump)	0.85
	fact	0.76	ridicare (pulling)	0.83
	fct	0.76	descarcare (offloading)	0.83
	fc	0.72	inchidere (closing)	0.82

Table 4: Samples of most similar words from Q/A word embeddings compared to Common Crawl. English translation is provided between parentheses.

made at the line level, each line from the original document is being marked for removal or to be kept. The removed lines include footer and header information from email exchanges, multiple email replies, tables dumped into text, tags, error messages, auto-replies, and sentences that did not have any connection to the problems stated in that request. We removed these categories and considered them irrelevant content. After this process, the pruned corpus shrunk to 73,000 words, aligned at the document and line level. We also decided to keep the email phrases, which are customary when starting and closing an email, as part of the content in order to later build heuristics around those to differentiate between multiple replies.

Based on the annotations we've made, the simplest approach to clean the corpus would be to create a binary classifier that can identify if a sentence or a group of sentences are to be removed or not.

Method	F_1	Accuracy
tf-idf classifier	0.746	0.890
emb classifier	0.714	0.873
tf-idf context proba	0.775	0.897
emb context proba	0.738	0.878
combined	0.774	0.893

Table 5: Cross-validation scores for different corpus cleanup methods.

This does not take into consideration the context or the surrounding sentences. We train a simple logistic regression classifier with regularization constant of 1, l2 penalty with liblinear solver (Fan et al., 2008; Pedregosa et al., 2011) on the tf-idf representations of each sentence. If a sentence has been removed by the annotator, it's a negative example, else it's a positive one. We compute the tf-idf for all tokens with diacritics removed from a sentence, including punctuation marks, numbers, function words, content words, and word bigrams. We carry a 5-fold cross-validation at the document level so that we don't shuffle the initial order of the sentences, obtaining an average cross-validation accuracy score of 0.89, and an average F_1 score of 0.74. Given the type of data at disposal, we were surprised to see such a good result, however, a closer look at the errors showed that the classifier was too rigid and biased towards the training data. When applied onto the entire corpus for cleanup, we could observe the removal of sentences and lines that should have been preserved. The source of this problem relies in the classifier not being aware of the context and surrounding sentences, and the tf-idf features being too dependent on the local training data to generalize well across the entire collection of texts that cover a wider diversity of topics than our annotations.

To overcome this overfitting problem we introduce two more variables: sentence probability in context and word embeddings. The first is used to reward sentences that have a small probability of being content by themselves, but have a high cumulative probability in the context of neighboring sentences. We establish a probability threshold (0.22) by grid search during cross validation. As for word embeddings, we used the previously trained models to create sentence representations from the word embeddings centroid of a sentence. We ignore function words and punctuation marks, and the words not present in the pretrained embedding model are set by default to vectors of zeros. Solely with this rudimentary sentence representations, we obtain a cross-validation classification accuracy of 0.87 and an average F_1 of 0.71,

slightly lower than the tf-idf representations. Table 5 contains the evaluation scores obtained during cross-validation. By combining the predictions of tf-idf models with the ones using embeddings, we obtain little improvements given the CV scores on the annotated dataset, however on the general dataset we observed a less restrictive behavior of the model that was able to preserve more easily out-of-domain content. Human evaluation is currently under way to asses the content quality of the selected sentences on subsamples from the larger dataset.

5 Conclusion

We provide a lexical comparative analysis of the language used in Q-A and Common Crawl corpora to the officially standardized one which is found in the dictionary. As a result of this study, we demonstrate that the actual use of language that prevails in the Q-A and Common Crawl corpora has a rather small overlap with the dictionary version (at most 60%). Moreover, in specific business communication circumstances, where the overlapping rate is expected to have increased values, code switching and non-standard expression are predominant, emphasizing once more the discrepancy between the day-to-day finacialized used language and the standardized one. In addition, we experiment with an approach to clean up the corpus based on a hybrid feature set consisting of word embeddings and tf-idf, to extract relevant content for further processing. Having few resources at disposal for Romanian, we believe it is mandatory to release parts of our data for reproducibility and future use.

Acknowledgments

This work has been supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS/CCCDI UEFISCDI, project number PN-III-P2-2.1-53BG/2016, within PNCDI III.

References

Gabriel Altmann and Michael H Schwibbe. 1989. *Das Menzerathsche Gesetz in Informationsverarbeitenden Systemen*. Georg Olms Verlag.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198. Association for Computational Linguistics.

Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social communication*, pages 343–359.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

F. Mosteller and L. D. Wallace. 1963. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309.

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. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

John Read. 2000. *Assessing vocabulary*. Cambridge University Press.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.