Evaluation of Embeddings in Medication Domain for Spanish Language Using Joint Natural Language Understanding

Surya Roca, Sophie Rosset, José García, Álvaro Alesanco

Evaluation of Embeddings in Medication Domain for Spanish Language Using Joint Natural Language Understanding

Surya Roca[1], Sophie Rosset[2], José García[1] and Álvaro Alesanco[1]

[1]*Aragón Institute of Engineering Research (I3A), University of Zaragoza, Zaragoza, Spain*

{surya, jogarmo, alesanco}@unizar.es

[2]*Université Paris-Saclay, CNRS, LIMSI, France*

sophie.rosset@limsi.fr

Abstract. Word embeddings have been widely used in Natural Language Processing as the input to neural networks. Such word embeddings can help in the understanding of the final objective and the keywords in a sentence. As such, in this work, we study the impact of different word embeddings trained with general and specific corpora using Joint Natural Language Understanding in a Spanish medication domain. We generate data using templates for training the model. The model is used for intent detection and slot-filling. We compare word2vec and fastText as word embeddings and ELMo and BERT as language models. We use three different corpora to train the embeddings: the training data generated for this scenario, the Spanish Wikipedia as general domain and the Spanish drug database as specialized data. The best result was obtained with word2vec continuous bag of words model learned with Spanish Wikipedia, obtaining a 71.77% F1-score for intent detection, an intent accuracy of 69.37% and a 74.36% F1-score for slot-filling.

Keywords. Intent Detection; Medication Management; Natural Language Understanding; Slot Filling; Word Embeddings

1. Introduction

Chatbots, or virtual assistants, are computer programs that interact with the users using text-based conversations. Nowadays, chatbots are used for numerous tasks, such as booking plane flights, ordering food and learning new languages. The benefits offered by such virtual assistants can significantly improve the effectiveness of healthcare services, especially when offered to elderly patients and children [1]. Patients can interactively obtain support, information or medical diagnosis by chatting with a virtual assistant. For some healthcare services like reminding the medicine intake, such chatbots can conveniently complement a health professional or personal caretaker. These healthcare chatbots can perform interesting tasks for users, but without a correct understanding of the circumstances under which the services are to be offered, patients may obtain wrong information or wrong monitoring. In other words, the chatbots should be intelligent enough to understand the scenario and the services desired by the users. Hence, the first and most important aim of the virtual assistant is to correctly understand the requests of the patients.

Natural Language Understanding (NLU) is a subtopic of Natural Language Processing, which is used to comprehend what an input text means and act consequently. NLU tries to predict the intention of the user and the keywords that the user is saying in a sentence. The tasks that the NLU performs are intent detection and slot-filling. Intent detection tries to label the sentence based on predefined intents [2]. Slot-filling can be defined as the action of tagging the words in a sentence with the slot types. Moreover, NLU systems have a learning approach that is based on the statistical representation of each word. This representation is obtained using embeddings. There exist two types of embeddings: static word embeddings (e.g. word2vec [3] or fastText [4]) which generates the same embedding for the same word regardless of the context, and contextualized word embeddings (a.k.a. embeddings from language models) (e.g. ELMo (Deep contextualized word representations) [5] or BERT (Bidirectional Encoder Representations from Transformers) [6]) which captures the word semantics in different contexts. Briefly, word embeddings are the representation of a word that is in a vocabulary into a vector of numerical values. The input of these embeddings is a text called corpus that serves to train the embedding. Corpus can be defined as a collection of structured texts used to do statistical analysis.

An interesting approach to studying the effect of the corpus in word embeddings was made by Wang et al. [7] where four different corpora in English were evaluated. Furthermore, Neuraz et al. [8] compared fastText with ELMo for different tasks in the clinical domain in French. Ghannay et al. [9] compared five different embeddings trained in English and French in five benchmark corpora for spoken language understanding. There are a few notable approaches with word embeddings in Spanish medical domain. Segura-Bedmar et al. [10] in their work, proposed an approach to simplify Drug Package Leaflets. Soares et al. [11] evaluated fastText in two different datasets. To the best of the authors' knowledge, a detailed study of different word embeddings and language models learned with different corpora for medication domain in Spanish has not been presented in literature before. Such a study can clearly reflect the true state-of- the art of NLU in medical chatbots and inspire more intelligent and advanced capabilities in medical chatbots.

As such, in this paper, we evaluate the impact of using general and specific corpora to learn different configurations of word embeddings. Furthermore, we test it in a medication domain, with data obtained from users. Finally, based on our evaluation, we establish the best configuration for our NLU system intended to provide medication management.

The rest of the paper is structured as follows. Section 2 provides a description of the system overview. Section 3 explains the methods used in the study. Section 4 presents and discusses the findings and results. Finally, Section 5 presents the conclusions of the paper.

2. Scenario Overview

In our previous work [12], we developed a virtual assistant for managing medication following a fixed interaction with options shown on a menu. After testing it with patients, we observed that a more natural interaction with the virtual assistant could improve the comfort of the users. Due to this necessity, we focused our work on adapting the virtual assistant, adding Natural Language Understanding (NLU) techniques.

In existing NLU techniques, there are different embedding configurations, corpora schemes and learning models. To find out the best configuration of embeddings and

corpora, we use different embeddings and datasets to obtain the best results for the medication domain. The different embeddings and datasets and the NLU system are shown in Fig. 1. In a medication domain, due to privacy issues, there is no access to corpus obtained from patients. Therefore, we have generated a total of 30,185 sentences based on templates and slot-filling for a medication scenario.

The proposed scenario uses the predictor multiLSTM [13]. This predictor uses bi-LSTM-CRF for joint NLU tasks (both slot-filling and intent detection). multiLSTM uses a corpus in IOB2 format to determine the intents and to detect the slots in a sentence. The original multiLSTM works with word2vec embeddings, training the word2vec with the generated data. We have adapted multiLSTM to be able to use fastText, ELMo and BERT as embeddings for the input of the neural network, using different embeddings trained with the proposed datasets for the comparison.

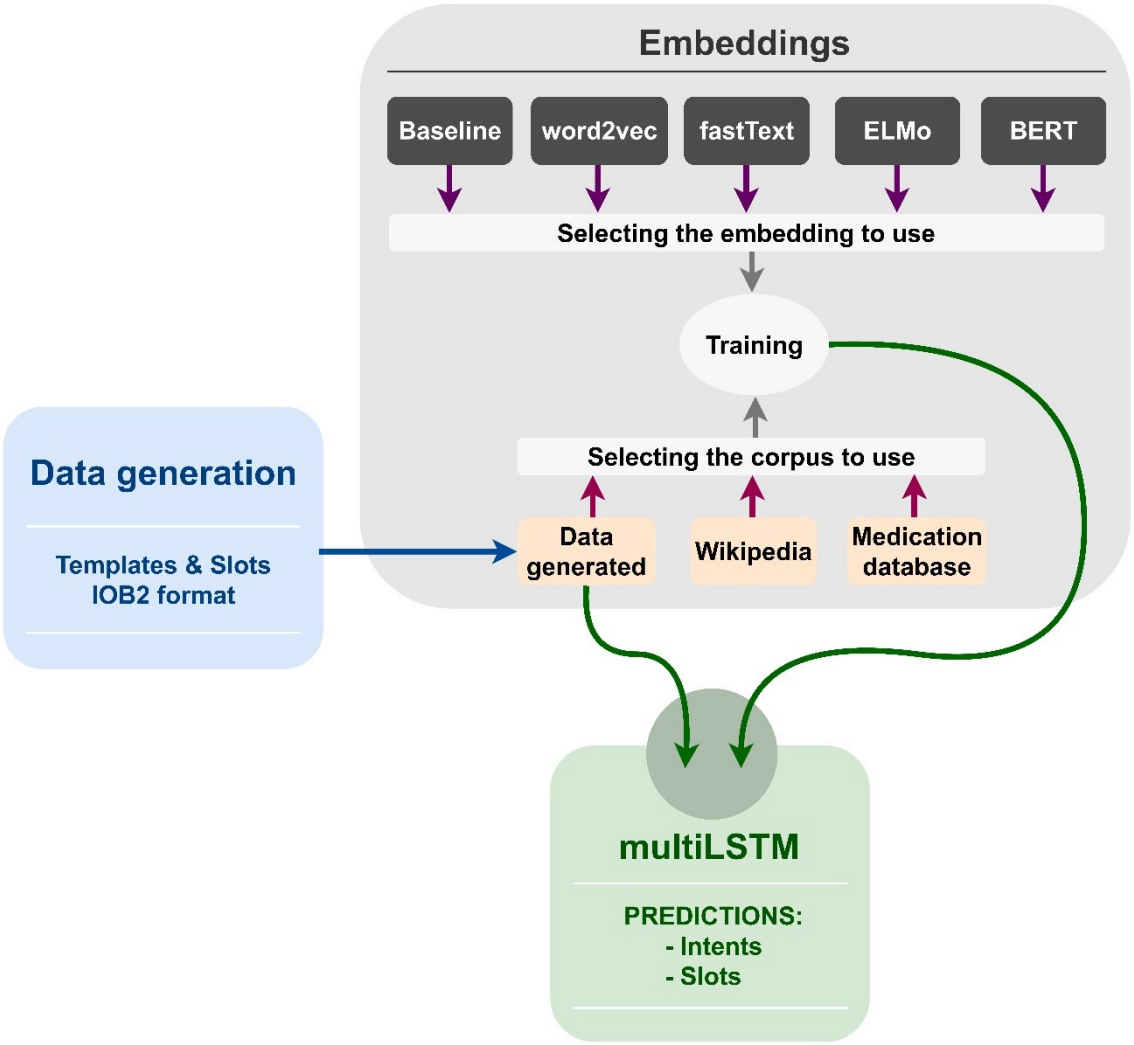

Fig. 1. Training scenario overview.

For our goal of comparing different embeddings and corpora to establish the best one, we have fixed the generated data and the model which predicts the intents and the slots. The study compares word2vec and fastText as word embeddings and ELMo and BERT as language models. The baseline is a word2vec implementation with

continuous skip-gram model which has been trained with only the training data generated for the medication scenario.

3. Methods

In this section, we explain in detail the methods we use to generate data and different embeddings, learning models and datasets used in the study.

3.1 Data generation for intent detection and slot filling

The data was generated using a set of 13,881 templates in total, for five different predefined intents. Also, we have defined six files, one for each slot type defined in our scenario. We have generated 20,000 sentences from these slots and templates for the training set (used during the learning process to fit the parameters of the model). Additionally, we have generated 5,000 sentences for the development set (used to evaluate the model while tuning the model's hyperparameters). Both generations was obtained using the algorithm proposed by Boulanger, described in [14]. We have further added additional sentences for seven basic intents, creating a total of 24,270 sentences for the training set and 5,915 sentences for the development set. These generated data have been used as an input to the multiLSTM, as well as a corpus to train the embeddings.

3.2 Training dataset

In this study, we have utilized three different training datasets for comparison. The first dataset is the data generated for training described in Section 3.1. The training data has a vocabulary with size 4,291 and with 274,114 train tokens. The second dataset was obtained from the Spanish Wikipedia. This dataset has a vocabulary with size 2,968,376 and with 683,501,282 train tokens. To be able to observe if a specific corpus in the domain has a positive impact on the results, we use the third dataset obtained from the Spanish medication database [15]. This dataset has a vocabulary with size 152,393 and with 143,741,382 train tokens. We have obtained the information pamphlet from 12,736 drugs and the technical datasheet from 12,813 drugs.

3.3 Embeddings

This study analyzes 13 different configurations of embeddings and corpora. As a baseline, we have a word2vec implementation with continuous skip-gram model of 300 dimensions trained with the generated training set. Also, we have a word2vec implementation with continuous bag of words (CBOW) model of 300 dimensions trained with the generated training set, the Wikipedia corpus, and the Spanish medication database. fastText embedding (using 300 dimensions and CBOW model) was trained with the generated training set, the Wikipedia corpus, and the Spanish medication database. ELMo embedding of 512 dimensions was trained with Wikipedia corpus and ELMo embedding of 1,024 dimensions was trained with the generated training set and Spanish medication database. Finally, we have BERT embedding of 768 dimensions trained with the generated training set, the Wikipedia corpus, and the Spanish medication database.

3.4 Test data

We have developed a chatbot with a specific conversation about patient medication management to collect the test data. We have obtained a total of 456 sentences from 14 people between 22 and 66 years old. We have manually filtered the sentences to obtain a final number of 382 sentences as test data. This filtering was made as not all

the sentences fit the proposed intents defined in our scenario of the medication domain. The test data has a vocabulary with size 291 and with 1,093 tokens. This vocabulary has 105 unknown words compared with the generated data for the training set. Comparing with Spanish Wikipedia, this vocabulary has 11 unknown words. Finally, this vocabulary has 34 unknown words compared with the Spanish medication database.

4. Results and Discussion

The results obtained from the simulations performed with the different configurations are shown in Table 1. We have used a weighted F1-score and an accuracy score to evaluate the results of the different configurations of embeddings and corpora. F1-score is defined by the equation 1, where precision is the division between the number of correct positive results (true positives, TP) and the number of all positive results obtained by the model (TP and false positives, FP). Recall is the division between the number of correct positive results (TP) and the number of all samples that should have been identified as positive (TP and false negatives, FN). The best value for F1-score is 1 and the worst is 0.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (1)$$

The best result was obtained with word2vec CBOW model trained with the Spanish Wikipedia, obtaining an F1-score of 74.36% for slot-filling, an F1-score of 71.77% for intent detection and an intent accuracy of 69.37%. The result of ELMo configuration trained with our generated data was obtained using the average of three layers and the results of ELMo configuration trained with Spanish Wikipedia and the Spanish medication database were obtained using the second LSTM hidden layer, as we observed that they had the best results. The best results for BERT configuration were obtained summing together the last four layers (each layer with 768 dimensions).

Table 1. Results for Joint NLU (in %).

Method	**Intent F1-score**	**Intent accuracy**	**Slot F1-score**
Baseline on generated data	69.08	67.02	64.81
Word2vec CBOW on generated data	60.03	57.33	58.95
Word2vec CBOW on Wiki	**71.77**	**69.37**	**74.36**
Word2vec CBOW on medication database	56.47	54.71	58.74
fastText on generated data	65.10	63.87	64.66
fastText on Wiki	68.29	68.85	72.92
fastText on medication database	59.59	58.12	56.60
ELMo on generated data	65.93	64.66	62.43
ELMo on Wiki	70.17	68.32	74.15
ELMo on medication database	65.41	65.71	62.50
BERT on generated data	66.57	64.92	70.55
BERT on Wiki	58.91	53.66	63.28
BERT on medication database	64.13	60.73	67.96

Comparing the results obtained with different training datasets, we observe the best results are obtained with the models that are trained with Spanish Wikipedia except for BERT. It seems that BERT has better performance when the training data is smaller and specific for the scenario. The worst results are obtained with the medication database in most of the embeddings. These results may be due to the fact that the medication database uses frequently a technical language, and in this context, the vocabulary used by the user with the virtual assistant is not very technical; usually the name of the medication is the only technical term in the conversation.

The main limitation of this model is the fact that our system predicts one of the defined intents and is not able to detect any other intent if the sentence is off-topic. Hence, as mentioned in Section 3.4, we need to filter the sentences obtained from the users to obtain only sentences that fit the intents of this scenario. This particular aspect of the model will be studied and enhanced further in future works.

5. Conclusions

The purpose of this study is to obtain the best result and discuss different configurations of embeddings with different corpora for a medication management scenario. Such comparative study can help in understanding the effectiveness of existing NLU techniques in medical domain and facilitate more advanced and intelligent features in chatbots. In our analysis, the NLU models in the medication domain show better results with the word2vec implementation with CBOW model, learned with the Spanish Wikipedia. This configuration can be considered as the best option to use in our virtual assistant. At this moment, we are working on adapting our virtual assistant with the NLU system obtained in this work.

Acknowledgements

Research funded by Ministerio de Economía, Industria y Competitividad from Gobierno de España and European Regional Development Fund (TIN2016-76770-R and BES-2017-082017) and Gobierno de Aragón and FEDER ``Construyendo Europa desde Aragón'' (T31_20R).

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Alesanco, Á., Sancho, J., Gilaberte, Y., Abarca, E., & García, J. (2017). Bots in messaging platforms, a new paradigm in healthcare delivery: application to custom prescription in dermatology. In *EMBEC & NBC 2017* (pp. 185-188). Springer, Singapore.
2. Stoica, A., Kadar, T., Lemnaru, C., Potolea, R., & Dînşoreanu, M. (2019, September). The impact of data challenges on intent detection and slot filling for the home assistant scenario. In *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)* (pp. 41-47). IEEE.
3. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

4. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.
5. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
6. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
7. Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., ... & Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, *87*, 12-20.
8. Neuraz, A., Looten, V., Rance, B., Daniel, N., Garcelon, N., Llanos, L. C., ... & Rosset, S. (2019). Do you need embeddings trained on a massive specialized corpus for your clinical natural language processing task?. In *MEDINFO 2019: Health and Wellbeing e-Networks for All* (pp. 1558-1559). IOS Press.
9. Ghannay, S., Neuraz, A., & Rosset, S. (2020, May). What is best for spoken language understanding: small but task-dependant embeddings or huge but out-of-domain embeddings?. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8114-8118). IEEE.
10. Segura-Bedmar, I., & Martínez, P. (2017). Simplifying drug package leaflets written in Spanish by using word embedding. *Journal of biomedical semantics*, *8*(1), 1-9.
11. Soares, F., Villegas, M., Gonzalez-Agirre, A., Krallinger, M., & Armengol-Estapé, J. (2019, June). Medical word embeddings for Spanish: Development and evaluation. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop* (pp. 124-133).
12. Roca, S., Hernández, M., Sancho, J., García, J., & Alesanco, Á. (2019, September). Virtual assistant prototype for managing medication using messaging platforms. In *Mediterranean Conference on Medical and Biological Engineering and Computing* (pp. 954-961). Springer, Cham.
13. multiLSTM. https://github.com/SNUDerek/multiLSTM. Accessed 27 Jan 2020
14. Boulanger, H. (2020, June). Évaluation systématique d'une méthode commune de génération. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 3: Rencontre des Étudiants Chercheurs en Informatique pour le TAL* (pp. 43-56). ATALA; AFCP.
15. Cima - centro de inforamción de medicamentos. https://cima.aemps.es/cima/publico/home.html. Accessed 24 Jan 2020