Data generation approaches for topic classification in multilingual spoken dialog systems

C. Montenegro
University of the Basque Country
(UPV/EHU)
San Sebastian, Spain
cae.montenegro@gmail.com

R. Santana
University of the Basque Country
(UPV/EHU)
San Sebastian, Spain
roberto.santana@ehu.eus

J. A. Lozano [ab]
[a]University of the Basque Country
(UPV/EHU)
San Sebastian, Spain
ja.lozano@ehu.eus

[b]Basque Center for Applied
Mathematics (BCAM)
Bilbao, Spain

ABSTRACT

The conception of spoken-dialog systems (SDS) usually faces the problem of extending or adapting the system to multiple languages. This implies the creation of modules specifically for the new languages, which is a time consuming process. In this paper, we propose two methods to reduce the time needed to extend the SDS to other languages. Our methods are particularly oriented to the topic classification and semantic tagging tasks and we evaluate their effectiveness on topic classification for three languages: English, Spanish, French.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; • **Computing methodologies** → **Neural networks**; **Machine learning algorithms**; • **Applied computing** → *Document management and text processing*.

KEYWORDS

spoken dialog systems, bilingual datasets, topic classification, neural networks

1 INTRODUCTION

Spoken-dialog systems (SDS) have become one of the most ubiquitous examples of human-machine interactions. In particular, they are very useful in assisting environments, where a natural and fluid communication between the user and the computer is desired. Development of open SDS, where the dialog is not structured

and can cover a variety of topics, usually requires a long time for development and important human resources.

This time can be significantly longer for multilingual SDS [1, 14, 16], understood as systems that are able to communicate with the user in different languages. Research on multilingual SDS has received considerable attention due to the emergence of commercial applications that can be deployed in many countries.

Extending or adapting a system that has been conceived for a particular language to other multiple languages is a difficult task. From an efficiency point of view, it makes sense to conceive strategies that allow us to reuse as many parts of the original design as possible. However, this can be complicated by the important interlingual differences and the unequal availability of corpora or other multilingual resources required to train the extensively used machine learning models.

In this paper, we focus on the natural language understanding (NLU) component of an open SDS designed to be the communication component of a virtual coach designed to improve the life of elderly [15]. The specific problem to solve is to create models which are able to perform topic classification of the user utterances in three different languages: English, Spanish and French. The rationale of incorporating topic classification as part of the NLU pipeline is to provide the user a richer conversational experience by switching the dialog manager to models specialized in each topic.

An added difficulty in our problem is that the availability of corpora for the three target languages is unequal. Therefore, we would like to create a modular system where information available in one language can be transferred or exploited at the time of learning models for the other languages.

Some works have proposed strategies for multilingual systems [4, 8, 10]. In this paper, we follow two different approaches. Both strategies are intended to label unlabeled dialog corpora that in turn are used to train topic-classification models. The first approach is based on the use of the Wordnet semantic network and synsets [18], which encapsulate information about different senses of commonly used words. In order to deal with training of the models for other languages, we identify synsets associated to English words that characterize topics, and then find aligned synsets in the other languages, which are used to build vocabularies. From these topic-specific vocabularies, unlabeled corpora in the target languages are labeled. In the second approach, we use parallel multilingual

corpora. We start from a topic specific vocabulary in the source language and then use it for identifying topic-specific utterances in the source corpus of a bilingual corpora. Then the corresponding utterances in the parallel target corpus are labeled with the same topic.

The paper is organized as follows. In the next section, we introduce the general SDS pipeline and explain the topic classification model. In Section 3, we present the two approaches to learn multilingual models. Related work is discussed in Section 4. Preliminary experiments on the comparison of the two methods are described in Section 5 together with the discussion of the results. The conclusions of our paper and topics for future research are presented in Section 6.
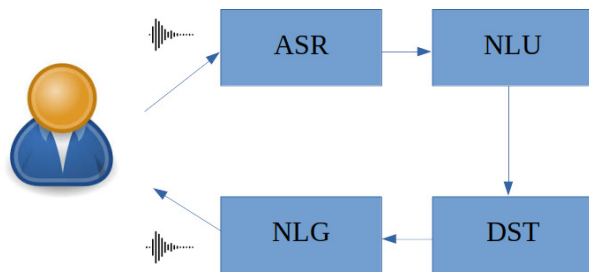
2 SDS PIPELINE

Figure 1: The architecture of an SDS

A dialogue system (DS), as the one illustrated in Figure 1, usually includes four main components (although there are many variations and this is mainly valid for a particular class of SDS):

(1) Automatic Speech Recognition (ASR) unit that transforms raw audio data into transcribed words.
(2) Natural Language Understanding (NLU) unit that gets the user utterances and usually transforms them to a semantic representation.
(3) A Dialogue State Tracking (DST) component that infers a current "state" of the dialogue from the semantics. Also, a Policy, based on the current state, decides what action to take, where an action can be to generate some semantics addressed to the user.
(4) Natural Language Generation (NLG) unit that, given the desired semantics, generates the text.

Although multilingual approaches can cover all the SDS components, we will focus on the NLU. Furthermore, dialogue systems can be classified into three classes [7]: 1) Task-oriented DS; 2) Domain-specific DS; and 3) Open-domain DS. We will constrain our analysis to the third class of SDS, which are usually called chatbots, and where the dialogue is not very well structured.

2.1 Topic classification with neural networks

In our NLU basic architecture illustrated in Figure 2, we can identify three main tasks, these are:

- End of turn detection: A conversation between two humans consists of a turn-taking transference of information, and replacing one human with a bot requires the detection of

the end of the user's end-of-turn pauses. The goal of an End of turn detection task is to detect this change of turn in a conversation between a human and the system in order to trigger the evaluation of the sentence or sentences received.
- Intent classification: The objective of this task is to identify the purpose of user's sentences, these purposes may be directives, wishes or questions, among others.
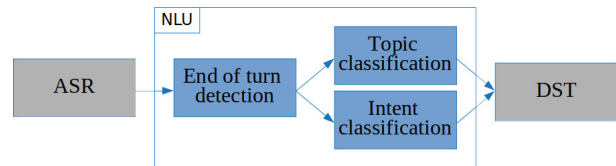- Topic classification: This tasks detects the subject of the conversation for each sentence.

Figure 2: The architecture of an NLU

As mentioned in Section 1, in order to select among the different topic models in the Dialog Manager, we need to detect each sentence's topic throughout a conversation. For this purpose, and since neural-network-based approaches to text classification have shown to be quite effective, we will use Long Short-term Memory (LSTM) recurrent networks [11] for this task. LSTM networks can be paired with an embedding layer that vectorizes each word. Word embedding is a distributed representation of words and greatly alleviates the data sparsity problem [2]. It is common to use pre-trained embedding models such as Glove [19] or Word2vec [17].

3 APPROACHES FOR EXTENDING MODELS TO OTHER LANGUAGES

3.1 Wordnet and synsets

Wordnet [18] is a dictionary where the main building blocks are the synsets (or word meanings). The dataset is organized by different types of semantic relations that relate words to synsets, and synsets among them. The meaning of a synset can be represented by a set of words that are used to express that meaning, and by a definition. Wordnet can be represented as a graph with different types of nodes and edges. Nodes represent synsets, and the edges the lexical and semantic relations between them. It was originally proposed for the Princeton WordNet (PWN) but since then different wordnets have been created for many languages[1].

One of the approaches used to construct WordNets in a given target language is the so-called *extend approach* [22] which consists of starting from a source WordNet (most frequently the PWN), translating a number of representative synsets into the target language and then enriching it by adding specific lexical relationships from that language. WordNets constructed in this way are easier to align, and more suitable for multilingual applications.

In this paper, we select WordNets of the target languages (Spanish and French). In these Wordnets there is a correspondence with the PWN synsets and therefore they are convenient for implementing multilingual SDS. As in previous work [3], it is assumed that if there are two synsets in PWN that hold a relation, then the same

---

[1]http://www.globalwordnet.org/gwa/wordnet_table.html

relation holds between the corresponding synsets in the target language

For the experiment proposed in Section 5, we generate two types of labelled corpora using topic-specific vocabularies that are created using synsets. The two strategies will be referred to "Wordnet approach" and "Parallel corpora".

3.2 Wordnet approach

We use WordNets to automatically create topic-labeled dialog databases in different languages. The key part of this approach is the manual identification of a number of synsets that serve to describe each topic, henceforth referred to *seed synsets*. Starting from these seed synsets, we can create, for each language, a set of words that serve to label the datasets. The pseudocode of the algorithm is shown in Algorithm 1.

Algorithm 1: **Wordnet approach**

1 For each topic, manually identify **seed synsets** associated to English words that characterize that topic.
2 Extend the **seed synsets** by computing the **closure set** that includes the hyperonymies of every word in the seed synsets.
3 Create a **topic word set** by joining all the words that have an associated synset in the closure set.
4 Starting from a dialog corpus, label each utterance as representative of a given topic if it contains at least k words in the corresponding topic word set.

The procedure shown in Algorithm 1 is done for every language we need. Although this algorithm generates a weak labelled dataset, increasing k parameter gives us more certainty that the sentence belongs to the automatically-assigned topic label.

3.3 Parallel corpora

This dataset-labelling method requires more specific starting resources. Firstly, we need to generate an English topic word set as described in Section 3.2. Moreover, the dialog to label with the English topic word set, must be translated to the desired second language. Although the translation might not be word by word, the meaning of the sentence should be the same, and therefore, the real topic label should also be the same in both languages.

The pseudocode of the algorithm is shown in Algorithm (2).

Algorithm 2: **Parallel corpora**

1 Generate English seed synsets for each topic.
2 Extend the seed synsets by computing the closure set that includes the hyperonymies of every word in the seed synsets.
3 Create a topic word set by joining all the words that have an associated synset in the closure set.
4 Starting from a dialog corpus, label each utterance as representative of a given topic if it contains at least k words in the corresponding topic word set.
5 Copy the labels of the English sentence to the **second language sentences**.

Table 1: LSTM architecture

Layer (type)	Output Shape	Param
lstm_1 (LSTM)	(None, 35)	127400
dense_1 (Dense)	(None, 1)	71

4 RELATED WORK

There are several works that use parallel texts to assist in the implementation of multilingual systems. Hwa et al [13] use an English parser, a word alignment package, and a large corpus of sentence-aligned bilingual text to induce Chinese dependency trees. Bentivogli and Pianta use word alignment between annotated source text and target text in a different language to transfer annotations [4]. Other works have used annotated parallel corpora for POS tagging across languages [8] and word sense disambiguation [10].

Great efforts have been made to create realistic dialog datasets. Our research benefits from these efforts. In particular, we have used the dataset recently introduced by Zhang et al. [23], to create the topic annotated dialog dataset.

The work presented in this paper has been constrained to the NLU component of the SDS. However, other approaches consider the multilingual characteristics, including all components of the model. Holzapfel et al [12] address the conception of multilingual SDS by combining vectorized context free grammars with inheritance, allowing the definition of general language independent rules and inherited rules for each language [12].

5 EXPERIMENTS

The main objective of our experiments is to make a preliminary validation of the models trained with datasets generated by the two described strategies, for the target languages considered. In this section, we first present the implementation and details of the models, and the procedure and metrics employed to estimate their accuracy. Then, we describe the different datasets used. Finally, we present the numerical results of the validation and discuss them.

5.1 LSTM models

LSTM models are one of the most widely used techniques, not only for topic classification, but for many text-related classification tasks(Section 2.1). Since the objective of these experiments is to validate the strategies we use to generate the labelled datasets, we will use a simple LSTM model, with one LSTM layer and the corresponding dense layer for outputs. The model architecture is described in Table 1.

Table 2 describes the general parameters and values used for training.

As mentioned in Section 2.1, embeddings are used to represent the text. We use the pre-trained embedding models available in Spacy[2] as described in Table 3.

---

[2] Available from http://www.spacy.com

Table 2: LSTM parameter values

Parameter	Value
n_batch	20
epochs	100
pad_sequence length	35
earlyStopping monitor	val_loss
earlyStopping patience	5

Table 3: Embeddings used for the different languages

Language	Model
English	en_core_web_lg
Spanish	es_core_news_md
French	fr_core_news_md

5.2 Experimental framework

The datasets we use in these experiments come from the open-subtitle[3] parallel corpora. Each dialogue appears in source and target languages. We use three bilingual datasets for the three languages:

- English: en-es.tmx (48.3M sentences)
- Spanish: en-es.tmx (48.3M sentences)
- French: en-fr.tmx (34.4M sentences)

These original datasets are not topic-labelled. Therefore, following the algorithm descriptions in Sections 3.2 and 3.3, we will generate two different labelled datasets starting from them. These datasets will be then used to train an LSTM model. To ensure that the model is learning to detect the topic, and not just keywords, we will randomly split the topic wordsets, using one of them for the learning process, and the other one for evaluation, as represented in Figure 3.

We have selected three topics (Nutrition, Leisure and Social) which are relevant in the specific approach to the creation of a virtual coach for elderly [15]. Examples of the seed synsets for each of these topics are the following:

- Nutrition: eat, food, breakfast, meal, restaurant
- Leisure: leisure, event, participant, spectator, sport, free time, hobbies
- Social: family, helper, care, nurse, assistant, doctor

Once we have created the TW-train and TW-test sets, generated from the synset seeds, for each topic of the language, we split each language's sentence database into train (50%), validation(25%), and test(25%) splits. Then we label train and validation splits following Algorithm 1 and using kewords in TW-train ($k = 2$). The test split is labelled with keywords in TW-test. We have, therefore, labelled train and test datasets for each language, which we will use for the approach experiment of Wordnet.

For Parallel corpora's experiment, we follow Algorithm 2 using the same labeling scheme previously described for the approach experiment of Wordnet. The specific characteristic in this experiment
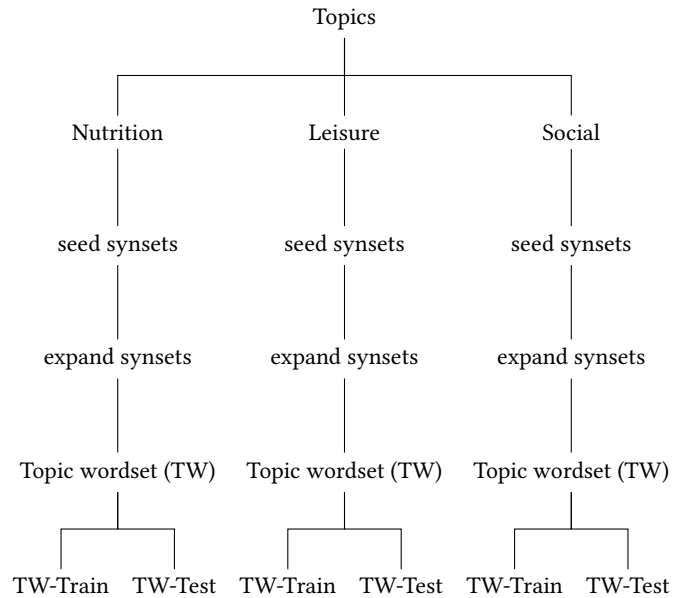
Figure 3: Topic wordset generation for experiments

Table 4: Size of the Topic wordsets by topic and language.

		Topic wordset sizes
English	**Nutrition**	1778
	Leisure	459
	Social	491
Spanish	**Nutrition**	300
	Leisure	110
	Social	99
French	**Nutrition**	720
	Leisure	190
	Social	230

is that the labels learned from the source language are also used to label the corresponding sentence in the target language.

5.3 Results

Table 4 shows the sizes of the wordsets organized by topic and language. An examination of the table reveals the diversity of the wordset sizes, not only regarding the topic, but also regarding the language. The sizes of these wordsets depend directly on the Wordnet model used for each language.

Tables 5 and 6 respectively show, for the parallel corpora and Wornet approaches, the sizes of train, validation and test sets used for training. An analysis of the tables shows that small wordsets have an effect on the final sizes of the train, validation and test sets for the Wordnet approach. However, a smaller wordset does not necessarily mean a poorer labeling of the sentences. Some examples of labeled sentences are shown in Table 7.

From the initially labeled sets, we learn the LSTM models. This experiment is repeated five times to account for the stochastic nature of the LSTM models. For each language, we evaluate the

Table 5: Sizes of the sets labeled with Parallel corpora approach.

		Train	Valid.	Test
English	Nutr.	10026	4972	11645
	Leisure	5160	2584	3226
	Social	34489	17225	5557
Spanish	Nutr.	9997	4935	11789
	Leisure	5154	2552	3298
	Social	34312	17225	5557
French	Nutr.	6674	3499	7858
	Leisure	3420	1789	2238
	Social	24190	12095	4184

Table 6: Sizes of sets labeled with Wordnet approach.

		Train	Valid.	Test
English	Nutr.	11955	6105	10848
	Leisure	2499	1226	5296
	Social	7341	3620	23159
Spanish	Nutr.	3156	1578	1899
	Leisure	563	262	169
	Social	501	265	803
French	Nutr.	10520	5151	4826
	Leisure	1009	511	398
	Social	2261	1110	5939

Table 7: Labeling examples from wordnet approach.

Sentence	Label asigned	Correct
i ll eat an apple and get back to you	Nutrition	yes
we can cut your leg off	Nutrition	no
swimming swimming in the swimming pool	Leisure	yes
and he s running round blaming me for the fire	Leisure	no
my mother she worked as a secretary for lord elms before she met my father	Social	yes
all you care about is the house the house the house	Social	no

accuracy of the models on the test datasets. As metrics to evaluate the performance of the model, we use the accuracy and two variants of the area under the curve (AUC) metric, the so-called AUC-micro and AUC-macro metrics. AUC is typically used in binary classification to study the output of a classifier. In order to extend AUC to multi-class or multi-label classification, it is necessary to binarize the output [9]. One ROC curve can be drawn per label, but it is also possible to draw a ROC curve by considering each element of the label indicator matrix as a binary prediction (micro-averaging) and then calculate the AUC-micro. The other AUC variant used for multi-class classification is the macro-averaging, which gives equal weight to the classification of each label in the AUC computation.

Table 8: Mean accuracies produced by the model in the five repetitions.

	Parallel corpora		Wordnet	
Language	Mean acc.	Std.	Mean acc.	Std.
English	54.9%	3.3%	66.3%	1.3%
Spanish	76.0%	3.7%	57.8%	1.1%
French	53.8%	3.3%	53.1%	1.4%

Table 9: Mean AUC-micro values for all models.

	Parallel corpora	Wordnet
English	71.9	82.4
Spanish	90.2	76.2
French	72.7	71.8

Table 10: Mean AUC-macro values for all models

	Parallel corpora	Wordnet
English	68.6	79.9
Spanish	87.5	75.8
French	72.7	70.3

Table 11: Confusion matrix for one of the trained networks in English.

	Prediction		
	Nutrition	Leisure	Social
Nutrition	42%	20%	37%
Leisure	22%	40%	37%
Social	15%	19%	65%

Table 8 shows the mean accuracies computed for the models for all languages and the two approaches evaluated. Tables 9 and 10 respectively show the AUC-micro and AUC-macro metrics. An analysis of the tables illustrates that the accuracies are well above chance levels for the two approaches and all languages. The results shown in tables 9 and 10 also reflect that the classifiers produce very accurate predictions. To futher illustrate the results of the classification process, the confusion matrix from one of the networks learnt can be seen in Table 11. The table shows that, in this example, the topics Leisure and Social are easier to classify than Nutrition. Finally, figures 4, 5 and 6 summarize a comparison of the performances of Wordnet and Parallel corpora approaches in terms of the accuracy.

5.4   Discussion

As represented in Table 7, the methods introduced in this paper generate noisy datasets. Nevertheless, these methods can easily generate customized datasets with specific labels that, depending on the problem, might not exist. The bigger the topic wordsets are for one language, the larger the labeled dataset will be. Therefore, the best method for a particular problem is the one that generates the largest dataset without modifying the k parameter or other criteria. This effect can be seen in figures 4, 5 and 6, where the
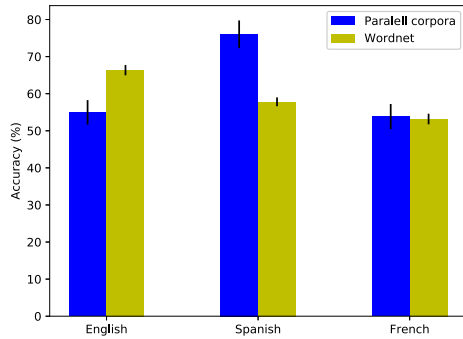
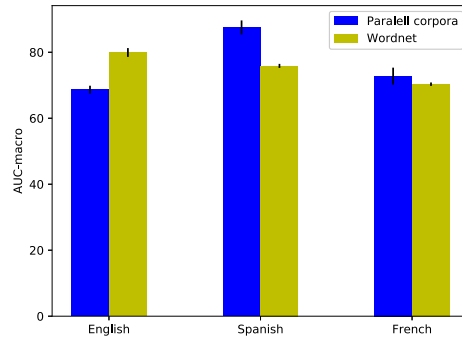**Figure 4: Accuracy of the Parallel corpora and Wordnet approaches for all languages.**
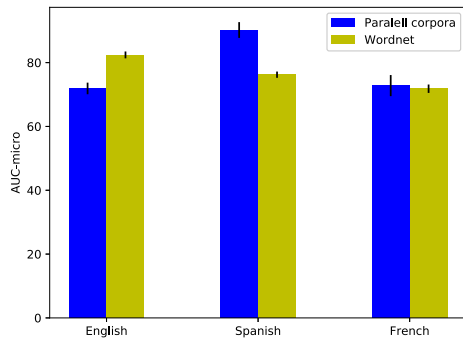

Figure 5: AUC-micro of the Parallel corpora and Wordnet approaches for all languages.

Spanish classification model increases greatly in performance when it is trained with the Parallel corpora dataset. The Parallel corpora dataset is formed by 51463 training sentences, while the Wordnet dataset is formed by only 4220 sentences.

The Parallel corpora approach has the advantage of not limiting the sentences to a specific set of words. This constraint is only on the English part of the process, but since the translations to the target languages are not constrained by a specific set of words, unlike the Wordnet approach, the vocabulary represented by these labeled sentences in the target language is richer.

6 CONCLUSIONS

While multilingual spoken dialog systems allow human-computer applications to reach a wider audience of individuals not necessarily familiar with computer applications, their implementation is time consuming and faces obstacles such as the lack of labelled datasets. In this paper, we have presented two approaches that, starting from a short description of relevant topics, allow the automatic labelling of datasets for training models. The assumption of our approach is that, even if the initial labelling of the datasets is imprecise,

Figure 6: AUC-macro of the Parallel corpora and Wordnet approaches for all languages.

the models will be able to capture the most common and most characteristic descriptors for each topic.

We have evaluated these approaches using parallel corpora. Our results indicate that high AUC values can be obtained using both methods. As in other machine learning applications, availability of data is a critical aspect to improve the results of the approach. The application of the algorithm could benefit from using larger bilingual corpora and more detailed Wordnet in the target languages. In this sense, there are many ongoing efforts on the extension of available WordNets and creation of new ones [5, 20, 21]. Also, the number of available multilingual corpora grows [6]. These works point to the possibility of extending our approach to include other languages.

The work described in this paper is a preliminary step for their incorporation as part of multilingual spoken dialog system conceived as a virtual coach for elderly people. The validation of the topic classification models as part of the whole SDS is another future direction of research.

7 ACKNOWLEDGMENTS

REFERENCES

[1] Jan Alexandersson, Norbert Reithinger, and Elisabeth Maier. 1997. Insights into the Dialogue Processing of VERBMOBIL. In *Proceedings of the fifth conference on Applied natural language processing*. Association for Computational Linguistics, 33–40.
[2] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3, Feb (2003), 1137–1155.
[3] Luisa Bentivogli and Emanuele Pianta. 2000. Looking for lexical gaps. In *Proceedings of the ninth EURALEX International Congress*. Stuttgart: Universität Stuttgart, 8–12.

[4] Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Natural Language Engineering* 11, 3 (2005), 247–261.

[5] Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. *Small* 8, 4 (2012), 5.

[6] Francis Bond, Shan Wang, Eshley Huini Gao, Hazel Shuwen Mok, and Jeanette Yi-wen Tan. 2013. Developing parallel sense-tagged corpora with wordnets. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse.* 149–158.

[7] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A Survey on Dialogue Systems: Recent Advances and New Frontiers. *CoRR* abs/1711.01731 (2017). http://arxiv.org/abs/1711.01731

[8] Elliott Franco Drábek and David Yarowsky. 2005. Induction of fine-grained part-of-speech taggers via classifier combination and crosslingual projection. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts.* Association for Computational Linguistics, 49–56.

[9] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.

[10] Alfio Massimiliano Gliozzo, Marcello Ranieri, and Carlo Strapparava. 2005. Cross-ing parallel corpora and multilingual lexical databases for WSD. In *International Conference on Intelligent Text Processing and Computational Linguistics.* Springer, 242–245.

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[12] Hartwig Holzapfel. 2005. Towards development of multilingual spoken dialogue systems. In *Proceedings of the 2nd Language and Technology Conference.*

[13] Rebecca Hwa, Philip Resnik, and Amy Weinberg. 2005. *Breaking the resource bot-tleneck for multilingual parsing.* Technical Report. Maryland University College Park. Institute for Advanced Computer Studies.

[14] Ivo Ipsic, Nikola Pavesic, France Mihelic, and Elmar Noth. 1999. Multilingual spoken dialog system. In *Proceedings of the IEEE International Symposium on Industrial Electronics, 1999. ISIE-99.*, Vol. 1. IEEE, 183–187.

[15] Asier López Zorrilla, Mikel de Velasco Vázquez, Jon Irastorza, Javier Mikel Olaso Fernández, Raquel Justo Blanco, and María Inés Torres Barañano. 2018. EMPATHIC: Empathic, Expressive, Advanced Virtual Coach to Improve Indepen-dent Healthy-Life-Years of the Elderly. *Procesamiento del Lenguaje Natural* 61 (2018), 167–170.

[16] Helen Meng, Shuk Fong Chan, Yee Fong Wong, Tien Ying Fung, Wai Ching Tsui, Tin Hang Lo, Cheong Chat Chan, Ke Chen, Lan Wang, Ting Yao Wu, et al. 2000. ISIS: A Multilingual Spoken Dialog System developed with CORBA and KQML agents. In *Sixth International Conference on Spoken Language Processing.*

[17] 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.* 3111–3119.

[18] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[19] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation. In *Empirical Methods in Natural Lan-guage Processing (EMNLP)*, Vol. 14. 1532–1543.

[20] Tommaso Petrolito and Francis Bond. 2014. A survey of wordnet annotated corpora. In *Proceedings of the Seventh Global WordNet Conference.* 236–245.

[21] Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multi-lingual resources. In *OntoLex.*

[22] Piek Vossen. 1998. A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers. doi* 10 (1998), 978–94.

[23] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? *CoRR* abs/1801.07243 (2018). http://arxiv.org/abs/1801.07243