Fully unsupervised low-dimensional representation of adverse drug reaction events through distributional semantics

Alicia Pérez, Arantza Casillas, Koldo Gojenola
IXA research group (http://ixa.si.ehu.eus)
University of the Basque Country (UPV-EHU)
{alicia.perez,arantza.casillas,koldo.gojenola}@ehu.eus

Abstract

Electronic health records show great variability since the same concept is often expressed with different terms, either scientific latin forms, common or lay variants and even vernacular naming. Deep learning enables distributional representation of terms in a vector-space, and therefore, related terms tend to be close in the vector space. Accordingly, embedding words through these vectors opens the way towards accounting for semantic relatedness through classical algebraic operations.

In this work we propose a simple though efficient unsupervised characterization of Adverse Drug Reactions (ADRs). This approach exploits the embedding representation of the terms involved in candidate ADR events, that is, drug-disease entity pairs. In brief, the ADRs are represented as vectors that link the drug with the disease in their context through a recursive additive model.

We discovered that a low-dimensional representation that makes use of the modulus and argument of the embedded representation of the ADR event shows correlation with the manually annotated class. Thus, it can be derived that this characterization results in to be beneficial for further classification tasks as predictive features.

1 Introduction

The aim of this work is to represent Adverse Drug Reactions (ADRs) efficiently so as to find drug related etiologies in Electronic Health Records (EHRs). Nebeker et al. (2004) defined an ADR as "a response to a drug which is noxious and which occurs as doses normally used". Finding ADRs efficiently is of much concern to pharmaco-surveillance and clinical documentation services. Personnel at pharmaco-surveillance services reads thousands of EHRs in order to detect this type of events and, furthermore, documentation services claim that, while ADRs should be reported by law, they seem to be under-reported (Dalianis et al., 2015).

From the natural language processing perspective, EHRs differ substantially from clinical literature as PubMed (Cohen and Demner-Fushman, 2014) in aspects such as syntax, the use of non-standard abbreviations (Okazaki et al., 2010; Kreuzthaler and Schulz, 2015), and misspellings (Dalianis, 2014). Within EHRs it is common to find the same concept expressed with different terms or surface-forms, synonyms or near-synonyms, either scientific latinised forms, common or lay variants or even vernacular naming, misspells, abbreviations etc. For example, we have found in the Spanish corpus we are working with a wide variety of ways to refer to the diagnostic term with code 600.00 from the ninth Clinical Modification of the International Classification of Diseases (WHO, 2014), namely, "benign prostatic hyperplasia" e.g. *hipertrofia benigna de próstata*, *hiperplasia BP*, *HBP-II*, *hiperplasia benigna de la prstata en estado II*, etc.

Distributional semantics has demonstrated to be a powerful approach to represent closely in a continuous space ($\mathbb{R}^n$) related entities. The rationale is to represent similar entities by means of close points in that space (or word-embedding) since close points render related meaning. Hence, embedding words

Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016), pages 50–59, Osaka, Japan, December 12th 2016.

through vectors opens the way towards semantic search (Thompson et al., 2016) as an alternative to the classical string-based methods that rely on keyword search. In this context, distributional semantics arises as a naturally appropriate framework as it exploits semantic similarity rather than the frequency of target surface-forms (Batista-Navarro et al., 2016). Deep learning (LeCun et al., 2015) has proven useful to model semantic relatedness based on corpus (Mikolov et al., 2013b). It accounts contextual information and patterns that occur in big amounts of unsupervised corpora to create sketches of word-forms that embed semantic relatedness. Mikolov et al. (2013a) proposed a distributed word embedding model that allowed to convey meaningful information on vectors derived from neural networks. With this approach, semantically related terms tend to be close in the vector space and the advantage is that closeness is a well studied metric in vector spaces (e.g. through Euclidean distance) while measuring semantic closeness is not trivial.

Our work is inspired by the example proposed by Mikolov et al. (2013a), where the authors present the association of the following terms: *Madrid* is to *Spain* what a *query* entity is to *France* or more formally as in (1).

$$\overrightarrow{Spain - Madrid} \approx \overrightarrow{France - query} \tag{1}$$

Here, each entity as *Madrid*, *Spain* or *France*, is represented as a point in a vector space that conveys meaningful information. The intuition is that close points correspond to semantically related entities. Accordingly, the hypothesis stands that similar vectors constructed by linking points also convey similar relations. Bearing this in mind and back to our domain, our research question is as follows: are drug-related aetiologies similarly represented in a semantic space? That is, given that a disease (e.g. "nosebleed") was caused by a drug (e.g. "sintrom") in a given EHR, can we extract other relations for their location in the vector space? Moreover, are similar the vectors that trigger ADR events and can be distinguished from those that do not form ADR events? We have tried to state this question formally through expression (2), where we denoted an ADR candidate by a disease-drug pair and, particularly, denoted as $\oplus$ the ADR events, $\ominus$ the non-ADR events and the sub-indices ($_i$ and $_j$) simply refer to a given particular instance in the data. We have tried to depict the research question through Figure 1 which shows relevant entities as points in a space and also a few ADR events through vectors linking drugs and diseases.

$$\begin{aligned} (\overrightarrow{Disease - Drug})_i^{\oplus} &\stackrel{?}{\approx} (\overrightarrow{Disease - Drug})_j^{\oplus} \\ (\overrightarrow{Disease - Drug})_i^{\oplus} &\stackrel{?}{\not\approx} (\overrightarrow{Disease - Drug})_j^{\ominus} \end{aligned} \tag{2}$$

The contribution of this paper is an efficient representation of ADR events with high correlation with the class ($\mathcal{C} = \{\ominus, \oplus\}$). The interest behind stands in its potential use for further supervised classification tasks as a stand-alone technique or, as it is our purpose, as predictive features for other classification techniques. All in all, we focus on representation while classification is out of the scope of this paper.

1.1 Related work

A big challenge of ADR event extraction is the fact that ADRs represent rare or infrequent events. In real EHRs we saw that ADRs represent 1% of the drug-disease pairs. That is, the ADR event extraction task is significantly skewed towards the negative class ($\ominus$ non-ADR events) in real EHRs and, hence, it represents a complex and still open problem for supervised classification. Regarding this issue, Henriksson (2015) created an artificially balanced corpus consisting of positive examples, health-care episodes coded with ADR-related diagnosis codes, and the negative examples were an equal number of randomly selected examples. By contrast, in this work we tackle the ADR extraction problem in its natural context, without avoiding the data skewness problem. Accordingly, we explore the scope of the proposed representation in the real context.

Last years authors have combined both supervised and unsupervised techniques to tackle entity recognition (Agerri and Rigau, 2016) and event extraction (Zhang et al., 2015; Zhou et al., 2015). Focusing on the clinical domain, the task presented by Henriksson et al. (2015) entailed the detection of health-care episodes that involved an ADR. They were pioneers in representing health-care episodes using semantic

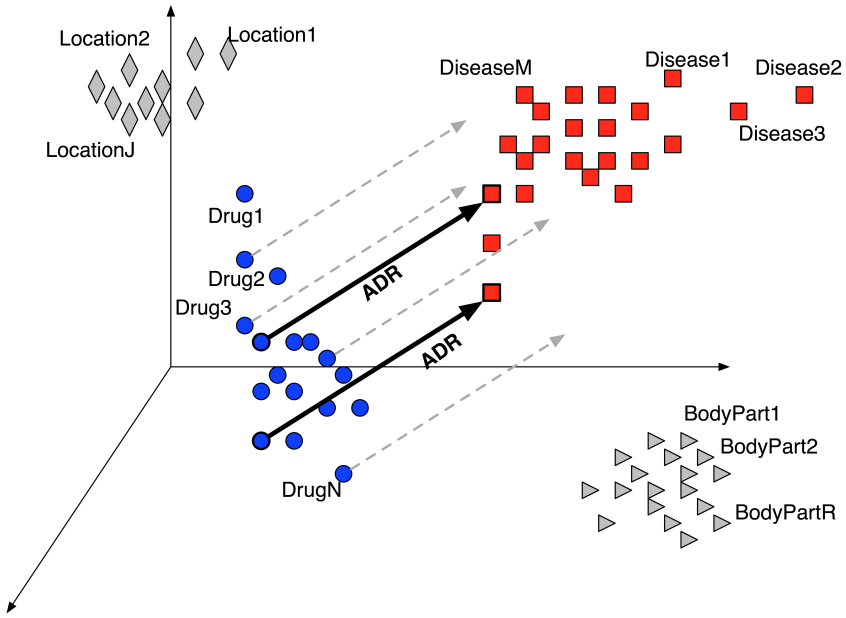

Figure 1: We represent the entities together with their context by means of word-embeddings (e.g. blue dots represent drugs, red squares diseases, triangles body-parts etc) and the ADR candidates through vectors linking a drug with a disease. The goal is to explore if the disposition of the ADR events is similar in the distributional semantic space.

spaces, that is, they extracted different representations of the documents and a semantic space was created for each component of the representation. The semantic spaces and other features were used as input to a machine learning algorithm. The task presented in (Henriksson et al., 2015) differs from ours since ADRs were a sub-set of diagnostic terms referred to as ADRs in the ICD-10-CM. This subset consisted of particular diseases that always had a drug as their cause. The goal was not to relate two entities (a drug and a disease) but, instead, it was to recognize a sub-set of disease-entities (a sub-set of ICDs). By contrast, we aim at extracting relations between drug and disease entities rather than subsets of entities.

Typically event extraction tasks and challenges (Pradhan et al., 2014) focus on the extraction of events that occur within the same sentence. Trying to relate entities that are in different sentences is by far much more complicated due to the amount of information that is required to take into account as the distance increases, not to mention anaphora and co-reference resolution. Nevertheless, the systems that only attempt at finding intra-sentence events might be discarding valuable information as many relations involve entities in different sentences. Indeed, in our set of EHRs, the inter-sentence events represent 51.7% of the positive instances, besides, on average the positive events are placed at a distance of 4 sentences but we found positive events involving entities further than 15 sentences. With the representation of each drug-disease pair proposed in this work we do not restrict ourselves to explore only intra-sentence events (as it tends to be the main trend in this area) but we also cope with inter-sentence events.

2 Methodology: ADRs as relation-vectors

Full classification systems rely entirely on predictive features to infer a model. The features represent, hence, a crucial source of knowledge. Our aim is to get an efficient representation of ADR events. By virtue of distributional semantics we prove that relevant features can be obtained. In an attempt to build a model able to extract events, the event should be characterized in an efficient manner. The key issue is that the representation itself should show correlation with the type of event (e.g. $\oplus$ for ADR events and $\ominus$ for non-ADR events).

We resort to semantic vector spaces to represent the entities (drugs and diseases), that is, each entity will be represented by its vector, calculated by the word2vec tool (Mikolov, 2016). As a result, each word has associated a point in an $\mathbb{R}^n$ vector-space. Nevertheless, a given drug does not always provoke

the same side effects, hence, the context turns out crucial to set the relations. Accordingly, we propose to encompass the entity together with its right and left contexts (being the contexts within a window of size m). Note that even though word vectors trained by means of Skip-gram model encode context information in the corpus, this context information reveals the global trend of the sample. With the general contextual trend, the model assigns a location in the space to each word. Nevertheless, the rationale behind the use of the context is that focusing on each EHR can leverage the local contextual information. In this regard, given an EHR in its textual form: $x^{(1)}, x^{(2)}, \cdots, x^{(i)}, \cdots, x^{(s)}$, and given that we identified $x^{(i)}$ as an entity (either drug or disease) that shall be taken into account as a candidate that could trigger an ADR event, the issue is how to render all the contextual information information as well. To do so, it is common practice to turn to the embedding of each word, that is, for a given word in the vocabulary, $x^{(k)} \in \Sigma$, get the corresponding embedding $\vec{\mathrm{x}}^{(k)} = (\mathrm{x}_1^{(k)}, \mathrm{x}_2^{(k)}, \cdots, \mathrm{x}_n^{(k)}) \in \mathbb{R}^n$ where $1 \leq k \leq |\Sigma|$ and n is the pre-fixed size for the dimension of the space. Nevertheless, in this work we do not rely only on their associated embedding but on a vector formed as a linear combination of the context vectors as stated through (3). In brief, for a given word $x^{(i)} \in \Sigma$, we get a vector $\vec{\mathrm{x}}^{(i)} \in \mathbb{R}^n$, but in such a way that it comprises not only the word but its context as well, that is, a window of m words to the left and to the right.

$$v(x^{(i-m)}, x^{(i-m+1)}, \cdots, x^{(i)}, \cdots, x^{(i+m-1)}, x^{(i+m)}) \quad = \quad \sum_{k=-m}^{m} \lambda_k \vec{\mathrm{x}}^{(i+k)} \tag{3}$$

Expression (3) represents a *recursive additive model* (Ferrone and Zanzotto, 2013) that makes a representation for each word within a given context, but adapted to the entity on the focus through a context of a given length (m). The weight λ_k balances the contribution of the context-word k in the representation, as an alternative to the *basic additive model* (Mitchell and Lapata, 2008; Zanzotto et al., 2010).

For instance, given that we would like to explore whether the disease $x^{(i)}$ was caused by the drug $y^{(j)}$ in a given EHR, first we represent the entities following the recursive additive representation. That is, for the word $x^{(i)}$ we get the vector $v(x^{(i-m)}, x^{(i-m+1)}, \cdots, x^{(i)}, \cdots, x^{(i+m-1)}, x^{(i+m)})$ that, for the sake of brevity, shall be referred as $\mathbf{x}$; likewise, for $y^{(j)}$ we get its contextual recursive representation $\mathbf{y}$. Given the representation in an n-dimensional semantic space of a disease, $\mathbf{x} \in \mathbb{R}^n$, and of a drug, $\mathbf{y} \in \mathbb{R}^n$, our goal is to carry out data analysis and measure if this characterization helps to represent the relatedness of those concepts and, thus, assess quantitatively if the semantic space helps to guess if the pair is an ADR event or not.

Quite naturally, we defined the contextual *relation vector* as the vector that starts in the point $\mathbf{x}$ and ends in $\mathbf{y}$ to represent the ADR. Note that $\overrightarrow{\mathbf{xy}} = \mathbf{y} - \mathbf{x}$, thus, $\overrightarrow{\mathbf{xy}} \in \mathbb{R}^n$. We explored both the *cosine similarity* between the entities $\mathbf{x}$ and $\mathbf{y}$ and also the *euclidean distance* between them or, what is equivalent, the argument and modulus of the relation vector $\overrightarrow{\mathbf{xy}}$. Cosine similarity is formally stated in (4) and the euclidean distance in (5), both of them are graphically depicted in Figure 2. It is well worth mentioning that x_i in (5) represents the i-th component of vector $\mathbf{x}$ and likewise, y_i for $\mathbf{y}$.

$$\cos(\theta) \quad = \quad \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}||\mathbf{y}|} \tag{4}$$

$$d(\mathbf{x}, \mathbf{y}) \quad = \quad |\overrightarrow{\mathbf{xy}}| = \left[\sum_{i=1}^{n} (x_i - y_i)^2 \right]^{\frac{1}{2}} \tag{5}$$

Note that the smaller the argument $\theta \equiv \angle(\mathbf{x}, \mathbf{y})$, the bigger its cosine and, thus, the more related the entities associated to $\mathbf{x}$ and $\mathbf{y}$. In the same way, the smaller the modulus $|\overrightarrow{\mathbf{xy}}|$, the more related the drug and disease entities. This is the reason for which we opted to provide the results in terms of the argument and the euclidean distance, both decrease as the entities get related.

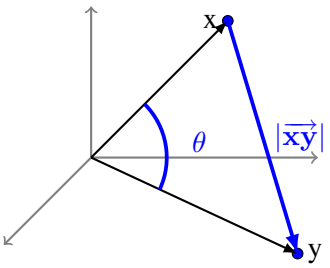

Figure 2: Argument θ, between vectors $\mathbf{x}$ and $\mathbf{y}$, and modulus of the relation vector $\mathbf{x}\vec{\mathbf{y}}$

Our aim is to analyze if the event-vectors $\overrightarrow{\mathbf{xy}}$ that relate ADRs show similarities (and likewise for non-ADRs). In other words, we aim at analyzing two sets of event-vectors that contain, respectively, ADR (or positive events) and non-ADR (or negative events), denoted as $\mathcal{S}_{\oplus}$ in (6) where $c(\overrightarrow{\mathbf{xy}})$ represents the real class of the event (equivalently, for $\mathcal{S}_{\ominus}$).

$$\mathcal{S}_{\oplus} = \{\overrightarrow{\mathbf{xy}} : c(\overrightarrow{\mathbf{xy}}) = \oplus\} \tag{6}$$

There are more sophisticated approaches that represent relation extraction through directed graphs that also make use of continuous vector spaces (Bordes et al., 2013). This kind of multi-relational data can represent entities as points in a vector space and relations as operations, such as projections, translations, etc. (Wang et al., 2014). In our work we explore a simple approach in a particular context where the relations are highly unbalanced. Nevertheless, relational machine learning approaches (Nickel et al., 2016) have demonstrated an for their ability to model and generalize relations. In particular, constraining the embedding in such a way that semantically related entities were placed in a lower-dimensional subspace (Jameel and Schockaert, 2016) seems applicable to our task. Provided that drug and disease entities would be represented in different subspaces while each sub-space would ensure that drug families would still be distinguished.

3 Experimental results

3.1 Task and corpus

We focus on EHRs written in Spanish by staff from the Galdakao-Usansolo and Basurto Hospitals. Admittedly, getting this kind of corpora to do research is not easy due to confidentiality issues (Cohen and Demner-Fushman, 2014), and it is even more difficult when it comes to explore other languages rather than English (Névéol et al., 2014). All in all, as Spanish is official in many countries, developing clinical text mining results for this language is of much interest, not only for the health systems but also for patients so that they get their EHRs in their own language.

The analysis of the proposed representation for relation vectors was built up based on an unsupervised or unannotated corpus, an in-domain medium-sized unsupervised set formed by 141,000 EHRs. From this partition we computed the word embeddings ($\vec{\mathbf{x}}^{(k)}$) that served to get the two features that are being proposed in this article (namely, θ and $|\overrightarrow{\mathbf{xy}}|$). Next, the assessment of the proposed representation was carried out on two independent supervised test sets not contained within the unsupervised set. In other words, we aimed to measure the correlation between θ and $|\overrightarrow{\mathbf{xy}}|$ with respect to the class ($\mathcal{C} = \{\oplus, \ominus\}$). The total number of tokens and documents involved in each set are shown in Table 1. To sum up, the

	tokens	docs	$\lvert\mathcal{S}_{\oplus}\rvert$	$\lvert\mathcal{S}_{\ominus}\rvert$
unsupervised	52×10^6	141,000	-	-
test-1	21×10^3	41	58	21,911
test-2	11×10^3	17	38	17,654

Table 1: Data sets: unsupervised to train the word embeddings and two supervised sets for testing.

resources exploited for the detection of ADRs are simply based on an unsupervised corpus, since we turn to word-embeddings and, from them, we derive the two proposed features (θ and $|\overrightarrow{\mathbf{xy}}|$). For the sake of curiosity, in this task we got an inter-annotator agreement of 82.86% on the test sets.

An inspection to Table 1 reveals a challenging (Monard and Batista, 2002; Phua et al., 2004; Mu et al., 2010) characteristic intrinsic of ADR detection: the classes are highly skewed being $\ominus$ the majority class. There are many works in this field that tackle ADR extraction with artificially balanced test sets (Henriksson, 2015). By contrast, we keep the repetition ratio as it is in the original sample of EHRs. Our aim is to check if this technique would help in real practice.

3.2 Results

The data analysis based on the proposed relation vector for ADR and non-ADR events is shown in Table 2. It presents the average argument θ and euclidean distance of relation vectors in each set ($\mathcal{S}_\oplus$ and $\mathcal{S}_\ominus$). Regarding the configuration of word2vec, we employed the skip-gram choice and a window of size s=5 requiring a n=300 dimensional vector space trained on the unsupervised set. Next, in order to represent each entity through the recursive additive model proposed in expression (3), a symmetric context of $m = 3$ tokens was chosen. Besides, we considered all the elements within the window equally influent and, hence, decided for $\lambda_i = 1$. Needless to say, this experimental setup involved a series of parameters (s, n, m, λ) that could have been fine-tuned on the basis of a supervised training corpus. Such a tuning would have helped to reassure the influence of each of them, for instance, with m an empirical comparison of the need to exploit local context and its scope when it comes to get the relation vector; with λ_i the influence of the context as the scope increases. While these empirical comparisons are of interest we found them out of the scope of this work.

	θ		$\|\overrightarrow{\mathbf{xy}}\|$	
	mean	stdev	mean	stdev
$\mathcal{S}_\oplus$	1.10	0.26	10.89	2.98
$\mathcal{S}_\ominus$	1.35	0.09	15.05	2.70

Table 2: Argument and modulus of ADR candidate events.

Figure 3 shows that ADR and non-ADR events from test-1 represented as relation vectors are statistically different in terms of both θ and $|\overrightarrow{\mathbf{xy}}|$. Hence, any of the proposed features (θ and $|\overrightarrow{\mathbf{xy}}|$) allows to distinguish between ADR and non-ADR events.

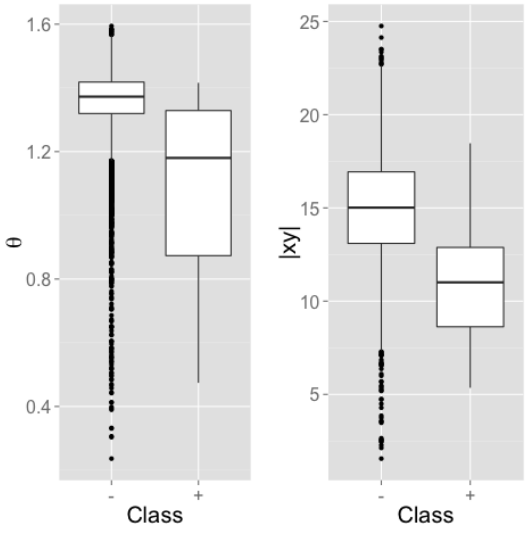

Figure 3: Box-plot for the argument, θ, and modulus, $|\overrightarrow{\mathbf{xy}}|$ with respect to the class of the event

The reader might wonder what would happen with an independent test from the same domain. We tackled this question through test-2 (presented in Table 1) and observed that the representation remains as stable as for the test-1. Indeed, both θ and $|\overrightarrow{\mathbf{xy}}|$ remained as in Table 2. Note that the model follows the intuition that related ADRs pose both small θ and small $|\overrightarrow{\mathbf{xy}}|$. In addition, one-way analysis of variance in the modulus stated that for this data-set, with respect to the null hypothesis of equal modulus, the p-value is 3.7×10^{-15}. The same applies to θ with a p-value of the same order of magnitude.

The proposed model is able to deal with inter-sentence and intra-sentence events. We wondered if the proposed event-vector representation gets degraded as the distance in sentences between the drug-disease pairs increases. In other words, does the event-vector remain stable in the vector-space despite they are in different sentences? Figure 4 comes to answer these questions. We explored the modulus of the relation vector for negative and positive instances. To be precise, we turned to the *relative location* of the drug $x^{(i)}$ with respect to the disease $y^{(j)}$ measured in sentences: $location(x^{(i)}, y^{(j)}) \equiv numSent(y^{(j)}) - numSent(x^{(i)})$. Whenever the drug and the disease are in the same sentence this location is 0; else, if the disease precedes the drug in the document, then the relative location is positive; otherwise, it is negative. Figure 4 depicts the box-plot associated to $|\vec{\mathbf{xy}}|$ for each class on three different location-ranges. We noticed that the regular way of reporting ADRs in EHRs in Spanish follows a scheme where the relative location is negative, and for them, $|\vec{\mathbf{xy}}|$ turned out very helpful to discriminate the class. As the trend changes, and particularly for very long positive relative locations, see the range "$(5, Inf]$" in Figure 4, the correlation of $|\vec{\mathbf{xy}}|$ with respect to the class decreases. We conclude that $|\vec{\mathbf{xy}}|$ is a helpful discriminant feature for ADR classification, particularly for those events that occur within the same sentence or relatively close (within 5 sentences), but also for the events that follow the trend and show a negative relative location despite of being far from each other.

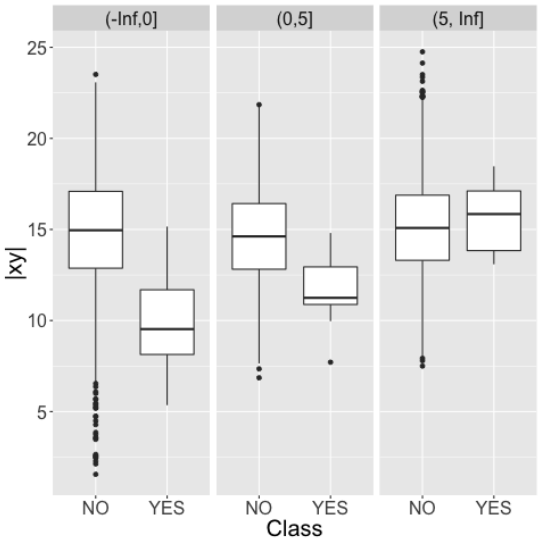

Figure 4: Impact on $|\vec{\mathbf{xy}}|$ of the relative location, in sentences, from the disease to the drug.

Even more, we carried out preliminary experiments with simple classifiers and, while it is out of the scope of this paper, the results are consistent with the hypotheses. The proposed features helped to discriminate ADR events. An example of the ADRs detected in a real EHR are shown in Figure 5 through Brat (Stenetorp et al., 2012). In this figure, the drug entities are marked in green with the tag `Grp_Medicamento` while the disease entities are marked in green with the tag `Grp_Enfermedad`. Positive ADR events are linked through arrows. In these examples all the ADRs occur within the same sentence, note that intra-sentence events.

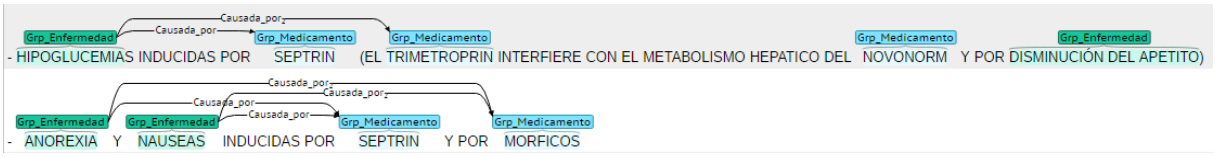

Figure 5: Example of ADRs detected in a real EHR.

In the mid-term we mean to test these features with a wide-suite of classifiers and focus, entirely, on classification techniques and results rather than on representation which is, indeed, the focus of this paper. Hence, we conclude that empirical observations lead us to support the representation introduced.

4 Concluding remarks and future work

In this work we proposed an efficient vector representation of drug-disease pairs, or ADR events, derived from distributional semantics. This representation of events showed clear correlation with respect to the class ($\oplus$ for ADR events and $\ominus$ for non-ADR events). In plain words, we made a linear combination of the context-vectors of each entity (either drug or disease) to get the vector representation of the entity within its context. Next, we formed the event by linking both entities yielding a relation vector.

Accordingly, we propose the use of a contextual recursive additive model to characterize each entity, either drug (represented as $\mathbf{x}$) or disease $\mathbf{y}$. Other approaches could have been explored, such as the mean vector (instead of the sum) of the context-vectors. Next, we related the drug trough the relation vector defined as $\overrightarrow{\mathbf{xy}}$. From here, we proved that two characteristics derived from this relation vector (θ and $\overrightarrow{\mathbf{xy}}$) showed clear correlation with respect to the class of the event ($\oplus$, $\ominus$). This work does not aim at proposing this low-dimensional representation as a stand-alone ADR event extraction technique, by contrast, we think of this as a prior step in order to leverage the representation of ADRs for subsequent supervised classification methods. This representation settled a basis for an ongoing work focused on the ADR classification.

Even though distributional semantics is known for its ability to embed related words in close positions of the vector space, there are still open challenges. Limitations of the approach explored in this paper stand on that we do not cope with ADRs expressed by means of aphoristic pronouns and co-referent expressions (such as “it”), even though distributional semantics could approach those terms for their co-appearance as well.

Future work is planned in two directions: on the one hand we aim at going ahead and try fully unsupervised classification of ADR events by improving this representation and enhancing it with LDA analysis; in parallel, we shall focus on feeding supervised classifiers with this approach and experimenting thoroughly if this low-dimensional vector representation can leverage the performance of current supervised methods.

Acknowledgments

The authors would like to thank the personnel of Pharmacy and Pharmacovigilance services of the Galdakao-Usansolo Hospital and personel of the Pharmacy service of the Basurto Hospital; also, Oier Lopez de Lacalle, it was for the reading group he is conducting at IXA that we approached first to deep learning; moreover, Josu Goikoetxea for his advice and helpful discussions on word embeedings; and not least, the anonymous reviewers for their constructive criticism. This work was partially funded by the Spanish Ministry of Science and Innovation (EXTRECM: TIN2013-46616-C2-1-R, TADEEP: TIN2015-70214-P) and the Basque Government (DETEAMI: Department of Health 2014111003).

References

Rodrigo Agerri and German Rigau. 2016. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63 – 82.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Riza Batista-Navarro, Jennifer Hammock, William Ulate, and Sophia Ananiadou. 2016. A text mining framework for accelerating the semantic curation of literature. In *International Conference on Theory and Practice of Digital Libraries*, pages 459–462. Springer.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, pages 2787–2795.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Kevin Bretonnel Cohen and Dina Demner-Fushman. 2014. *Biomedical Natural Language Processing*. Natural Language Processing. John Benjamins Publishing Company.

Hercules Dalianis, Aron Henriksson, Maria Kvist, Sumithra Velupillai, and Rebecka Weegar. 2015. Health bank - a workbench for data science applications in healthcar. In J. Krogstie, G. Juel-Skielse, and V. Kabilan, editors, *Proceedings of the CAiSE-2015 Industry Track co-located with 27th Conference on Advanced Information Systems Engineering*, volume 1381, pages 1–18.

Hercules Dalianis. 2014. Clinical text retrieval-an overview of basic building blocks and applications. In *Professional Search in the Modern World*, pages 147–165. Springer.

Lorenzo Ferrone and Fabio Massimo Zanzotto. 2013. Linear compositional distributional semantics and structural kernels. In *Joint Symposium on Semantic Processing.*, page 85. Citeseer.

Association for Computing Machinery. 1983. In *Computing Reviews*, volume 24, pages 503–512.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Aron Henriksson, Jing Zhao, Henrik Bostrom, and Hercules Dalianis. 2015. Modeling electronic health records in ensembles of semantic spaces for adverse drug event detection. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pages 343–350. IEEE.

Aron Henriksson. 2015. Representing clinical notes for adverse drug event detection. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 152–158, Lisbon, Portugal, September. Association for Computational Linguistics.

Shoaib Jameel and Steven Schockaert. 2016. Entity embeddings with conceptual subspaces as a basis for plausible reasoning. In *European Conference on Artificial Intelligence*, volume 22, pages 1353–1361.

Markus Kreuzthaler and Stefan Schulz. 2015. Detection of sentence boundaries and abbreviations in clinical narratives. *BMC medical informatics and decision making*, 15(Suppl 2):S4.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

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

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Tomas Mikolov. 2016. word2vec: Tool for computing continuous distributed representations of words. Accessed 2016-07-08, https://code.google.com/p/word2vec/.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *ACL*, pages 236–244.

Maria Carolina Monard and Gustavo EAPA Batista. 2002. Learmng with skewed class distrihutions. *Advances in Logic, Artificial Intelligence, and Robotics: LAPTEC 2002*, 85:173.

Tingting Mu, Xinglong Wang, Jun'ichi Tsujii, and Sophia Ananiadou. 2010. Imbalanced classification using dictionary-based prototypes and hierarchical decision rules for entity sense disambiguation. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 851–859. Association for Computational Linguistics.

Jonathan R. Nebeker, Paul Barach, and Matthew H. Samore. 2004. Clarifying adverse drug events: A clinician's guide to terminology, documentation, and reporting. *Annals of Internal Medicine*, 140(10):795–801.

Aurélie Névéol, Hercules Dalianis, Guergana Savo va, and Pierre Zweigenbaum. 2014. Panel: Clinical natural language processing in languages oth er than english. In *American Medical Informatics Association (AMIA)*.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33.

Naoaki Okazaki, Sophia Ananiadou, and Jun'ichi Tsujii. 2010. Building a high-quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*, 26(9):1246–1253.

Clifton Phua, Damminda Alahakoon, and Vincent Lee. 2004. Minority report in fraud detection: classification of skewed data. *Acm sigkdd explorations newsletter*, 6(1):50–59.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. Semeval-2014 task 7: Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

Paul Thompson, Riza Theresa Batista-Navarro, Georgios Kontonatsios, Jacob Carter, Elizabeth Toon, John McNaught, Carsten Timmermann, Michael Worboys, and Sophia Ananiadou. 2016. Text mining the history of medicine. *PloS one*, 11(1):e0144717.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, pages 1112–1119. Citeseer.

WHO. 2014. International classification of diseases (ICD). World Health Organization.

Fabio Massimo Zanzotto, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1263–1271. Association for Computational Linguistics.

Congle Zhang, Stephen Soderland, and Daniel Weld. 2015. Exploiting parallel news streams for unsupervised event extraction. *Transactions of the Association for Computational Linguistics*, 3:117–129.

Deyu Zhou, Liangyu Chen, and Yulan He. 2015. An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization.