Exploration of Adverse Drug Reactions in Semantic Vector Space Models of Clinical Text

Aron Henriksson ARONHEN@DSV.SU.SE
Department of Computer and Systems Sciences (DSV), Stockholm University, Sweden

Maria Kvist MARIA.KVIST@KAROLINSKA.SE
Department of Clinical Immunology and Transfusion Medicine, Karolinska University Hospital, Sweden
Department of Computer and Systems Sciences (DSV), Stockholm University, Sweden

Martin Hassel XMARTIN@DSV.SU.SE
Department of Computer and Systems Sciences (DSV), Stockholm University, Sweden

Hercules Dalianis HERCULES@DSV.SU.SE
Department of Computer and Systems Sciences (DSV), Stockholm University, Sweden

Abstract

A novel method for identifying potential side-effects to medications through large-scale analysis of clinical data is here introduced and evaluated. By calculating distributional similarities for medication-symptom pairs based on co-occurrence information in a large clinical corpus, many known adverse drug reactions are successfully identified. These preliminary results suggest that semantic vector space models of clinical text could also be used to generate hypotheses about potentially unknown adverse drug reactions. In the best model, 50% of the terms in a list of twenty are considered to be conceivable side-effects. Among the medication-symptom pairs, however, diagnostic indications and terms related to the medication in other ways also appear. These relations need to be distinguished in a more refined method for detecting adverse drug reactions.

1. Introduction

The prevalence of adverse drug reactions (ADRs) constitutes a major public health issue. In Sweden it has been identified as the seventh most common cause of death (Wester et al., 2008). The prospect of being able to detect potentially unknown or undocumented side-effects of medicinal substances by automatic means thus comes with great economic and health-related benefits.

As clinical trials are generally not sufficiently extensive to identify all possible side-effects – especially less common ones – many ADRs are unknown when a new drug enters the market. Pharmacovigilance is therefore carried out throughout the life cycle of a pharmaceutical product and is typically supported by reporting systems and medical chart reviews (Chazard, 2011). Recently, with the increasing adoption of Electronic Health Records (EHRs), several attempts have been made to facilitate this process by detecting ADRs from large amounts of clinical data. Many of the previous attempts have focused primarily on structured patient data, thereby missing out on potentially relevant information only available in the narrative parts of the EHRs.

The aim of this preliminary study is to investigate the application of distributional semantics, in the form of Random Indexing, to large amounts of clinical text in order to extract drug-symptom pairs. The models are evaluated for their ability to detect known and potentially unknown ADRs.

2. Background

Extracting ADRs automatically from large amounts of data essentially entails discovering a relationship – ideally a *cause-and-effect* relationship – between biomed-

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

ical concepts, typically between a medication and a symptom/disorder. A potentially valuable source for this are EHRs, in which symptoms – sometimes as possible adverse reactions to a drug – are documented at medical consultations of various kinds. In many cases, however, the prescription of a named medication precedes the description of symptoms (diagnostic indications): this obviously does not correspond to the required temporal order of a cause-and-effect relationship for side-effects. In other cases, the intake of a medication precedes the appearance of new symptoms, which could thus be side-effects or merely due to a second disease appearing independently from earlier medical conditions. The different ways in which drugs and symptoms can be documented in EHRs present a serious challenge for automatic detection of ADRs.

2.1. Related Research

Attempts to discover a relationship between drugs and potential side-effects are usually based on co-occurrence statistics, semantic interpretation and machine learning (Doğan et al., 2011). For instance, Benton et al. (2011) try to mine potential ADRs from online message boards by calculating co-occurrences with drugs in a twenty-token window.

The increasing digitalization of health records has enabled the application of data mining – sometimes with the incorporation of natural language processing – to clinical data in order to detect Adverse Drug Events (ADEs) automatically. Chazard (2011) mines French and Danish EHRs in order to extract ADE detection rules. Decision trees and association rules are employed to mine structured patient data, such as diagnosis codes and lab results; free-text discharge letters are only mined when ATC[1] codes are missing. Wang et al. (2010) first map clinical entities from discharge summaries to UMLS[2] codes and then calculate co-occurrences with drugs. Arakami et al. (2010) assume a similar two-step approach, but use Conditional Random Fields to identify terms; to identify relations, both machine learning and a rule-based method are used. They estimate that approximately eight percent of their Japanese EHRs contain ADE information. In order to facilitate machine learning efforts for automatic ADE detection, MEDLINE[3] case reports are annotated for mentions of drugs, adverse effects, dosages, as well as the relationships between them. The *ADE* corpus is now freely available (Gurulingappa et al., 2012).

2.2. Distributional Semantics

Another way of discovering relationships between concepts is to apply models of distributional semantics to a large text collection. Common to the various models that exist is their attempt to capture the meaning of words based on their distribution in a corpus of unannotated text (Cohen & Widdows, 2009). These models are based on the *distributional hypothesis* (Harris, 1954), which states that words with similar distribution in language have similar meanings. Traditionally the semantic representations have been spatial or geometric in nature by modeling terms as vectors in high-dimensional space, according to which contexts – and with what frequency – they occur in. The semantic similarity of two terms can then be quantified by comparing their distributional profiles.

Random Indexing (see e.g. Sahlgren 2006) is a computationally efficient and scalable method that utilizes word co-occurrence information. The meaning of words are represented as vectors, which are points in a high-dimensional vector space that, for each observation, move around so that words that appear in similar contexts (i.e. co-occur with the same words) gravitate towards each other. The distributional similarity between two terms can be calculated by e.g. taking the cosine of the angles between their vectorial representations.

3. Method

The data from which the models are induced is extracted from the Stockholm EPR Corpus (Dalianis et al., 2009), which contains EHRs written in Swedish[4]. A document comprises clinical notes from a single patient visit. A patient visit is difficult to delineate; here it is simply defined according to the continuity of the documentation process: all free-text entries written on consecutive days are concatenated into one document. Two data sets are created: one in which patients from all age groups are included (20M tokens) and another where records for patients over fifty years are discarded (10M tokens). This is done in order to investigate whether it is easier to identify ADRs caused in patients less likely to have a multitude of health issues. The web of cause and effect may

[1] Anatomical Therapeutic Chemical, used for the classification of drugs.

[2] Unified Medical Language System: `http://www.nlm.nih.gov/research/umls/`.

[3] Contains journal citations and abstracts for biomedical literature: `http://www.nlm.nih.gov/pubs/factsheets/medline.html`.

[4] This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2009/1742-31/5.

thereby be disentangled to some extent. Preprocessing is done on the data sets in the form of lemmatization and stop-word removal.

Random Indexing is applied on the two data sets with two sets of parameters, yielding a total of four models. A model is induced from the data in the following way. Each unique term in the corpus is assigned a *context vector* and a *word vector* with the same predefined dimensionality, which in this case is set to 1,000. The *context vectors* are static, consisting of zeros and a small number of randomly placed 1s and -1s, which ensures that they are nearly orthogonal. The *word vectors* – initially empty – are incrementally built up by adding the *context vectors* of the surrounding words within a sliding window. The context is the only model parameter we experiment with, using a sliding window of eight (4+4) or 24 (12+12) surrounding words.

The models are then used to generate lists of twenty distributionally similar terms for twenty common medications that have multiple known side-effects. To enable a comparison of the models induced from the two data sets, two groups of medications are selected: (1) drugs for cardiovascular disorders and (2) drugs that are not disproportionally prescribed to patients in older age groups, e.g. drugs for epilepsy, diabetes mellitus, infections and allergies. Moreover, as side-effects are manifested as either symptoms or disorders, a vocabulary of unigram SNOMED CT[5] terms belonging to the semantic categories *finding* and *disorder* is compiled and used to filter the lists generated by the models. The drug-symptom/disorder pairs are here manually evaluated by a physician using lists of indications and known side-effects[6].

4. Results

Approximately half of the model-generated suggestions are disorders/symptoms that are conceivable side-effects; around 10% are known and documented (Table 1). A fair share of the terms are indications for prescribing a certain drug (∼10-15%), while differential diagnoses (i.e. alternative indications) also appear with some regularity (∼5-6%). Interestingly, terms that explicitly indicate mention of possible adverse drug events, such as *side-effect*, show up fairly often. However, over 20% of the terms proposed by all of the models are obviously incorrect; in some cases they are not symptoms/disorders, and sometimes they are simply not conceivable side-effects.

Table 1. Results for models built on the two data sets (all age groups vs. only ≤ 50 years) with a sliding window context of 8 (SW 8) and 24 words (SW 24) respectively. S-E = side-effect; D/S = disorder/symptom.

	All		≤ 50	
Type	SW 8	SW 24	SW 8	SW 24
Known S-E	11	11	10	11
Potential S-E	37	39	35	35
Indication	14	14	10	10
Alt. indication	6	6	5	6
S-E term	10	9	7	7
D/S, not S-E	15	14	23	21
Not D/S	7	7	10	10
SUM %	100	100	100	100

Using the model induced from the larger data set – comprising all age groups – slightly more potential side-effects are identified, whereas the proportion of known side-effects remains constant across models. The number of incorrect suggestions increases when using the smaller data set. The difference in the scope of the context, however, does not seem to have a significant impact on these results. Similarly, when analyzing the results for the two groups of medications separately and vis-à-vis the two data sets, no major differences are found.

5. Discussion

The models produce some interesting results, being able to identify indications for a given drug, as well as known and potentially unknown ADRs. In a more refined method for identifying unknown side-effects, indications and known side-effects could easily be filtered out automatically. A large portion of the suggested terms were, however, clearly neither indications nor drug reactions. One reason is that the list of SNOMED CT *findings* and *disorders* contains several terms that are neither symptoms nor disorders. Refining this list is an obvious remedy, which could be done by for instance using a list of known side-effects. We also found that many of the unexpected words were very rare in the data, which means that their semantic representation is not statistically well-grounded: such words should probably be removed. Many common and expected side-effects did, on the other hand, not show up, e.g. *headache*. This could be due to the prevalence of this term in many different contexts, which diffuses its meaning in such a way that it is not distributionally similar to any particular other term. Moreover, since the vocabulary list was not lemmatized, but the health records were, some terms that were not already in their base form could not be proposed by the models. Some

[5] **S**ystematized **No**menclature of **Med**icine – **C**linical **T**erms: http://www.ihtsdo.org/snomed-ct/.

[6] http://www.fass.se, a medicinal database.

known side-effects were not present in the vocabulary list.

A limitation of these models is that they are presently restricted to unigrams; however, many side-effects are multiword expressions. Moreover, the models in these experiments were induced from data comprising all types of clinical notes; however, Wang et al. (2010) showed that their results improved by using specific sections of the clinical narratives. It would also be interesting to generate drug-symptom pairs for multiple drugs that belong to the same ATC code and extract potential side-effects common to them both. This could be a means to find stronger suspicions of ADRs related to a particular chemical substance.

6. Conclusion

By calculating distributional similarities for medication-symptom pairs based on co-occurrence information in a large clinical corpus, many known ADRs are successfully detected. These preliminary results suggest that semantic vector space models of clinical text could also be used to generate hypotheses about potentially unknown ADRs. Although several limitations must be addressed for this method to demonstrate its true potential, distributional similarity is a useful notion to incorporate in a more sophisticated method for detecting ADRs from clinical data.

Acknowledgments

This work was supported by the Swedish Foundation for Strategic Research through the project *High-Performance Data Mining for Drug Effect Detection* (ref. no. IIS11-0053) at Stockholm University, Sweden.

References

Arakami, E., Miura, Y., Tonoike, M., Ohkhuma, T., Masuichi, H., Waki, K., and Ohe, K. Extraction of adverse drug effects from clinical records. In *MEDINFO 2010*, pp. 739–743. IOS Press, 2010.

Benton, A., Ungar, L., Hill, S., Hennessy, S., Mao, J., Chung, A., Leonard, C.E., and Holmes, J.H. Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *Journal of Biomedical Informatics*, 44:989–996, 2011.

Chazard, Emmanuel. *Automated Detection of Adverse Drug Events by Data Mining of Electronic Health Records*. PhD thesis, Universite Lille Nord de France, 2011.

Cohen, T. and Widdows, D. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42:390–405, 2009.

Dalianis, H., Hassel, M., and Velupillai, S. The Stockholm EPR Corpus: Characteristics and Some Initial Findings. In *Proc. of ISHIMR*, pp. 243–249, 2009.

Doğan, R. I., Névéol, A., and Lu, Z. A context-blocks model for identifying clinical relationships in patient records. *BMC Bioinformatics*, 12(3):1–11, 2011.

Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., and Toldo, L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, http://dx.doi.org/10.1016/j.jbi.2012.04.008, 2012.

Harris, Z. S. Distributional structure. *Word*, 10:146–162, 1954.

Sahlgren, M. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Stockholm University, 2006.

Wang, X., Chase, H., Markatou, M., Hripcsak, G., and Friedman, C. Selecting information in electronic health records for knowledge acquisition. *Journal of Biomedical Informatics*, 43(4):595–601, 2010.

Wester, K., Jönsson, A. K., Spigset, O., Druid, H., and Hägg, S. Incidence of fatal adverse drug reactions: a population based study. *British Journal of Clinical Pharmacology*, 65(4):573–579, 2008.