A Statistical Method for Learning the Residual Quickly, Accurately, and Language Independently

Anni Coden[1], Daniel Gruhl[2], Neal Lewis[2], Michael Tanenblatt[1]
{anni,dgruhl,nrlewis,mtan}@us.ibm.com
[1] IBM, T.J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532
[2] IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120

Mining knowledge from various data sources is rapidly increasing. A critical first step in this process is to determining semantically equivalent concepts, in effect creating and maintaining up-to-date semantic lexicons. In the medical domain, many concepts must be gleaned from unstructured clinical text, whose format, language, abbreviations and punctuation may vary within and across institutions, as well as by specialty.

We first focused on the identification of drug names due to the importance of mitigating adverse drug reactions, interactions, and misuse. Although terminologies and ontologies (herein collectively referred to as dictionaries) exist, simple lookup is insufficient: dictionaries are incomplete, there is an ever increasing number of trade names, standardized dictionary entries often differ from how terms are used in clinical texts, there are various (mis)spellings, as well as over-the-counter drugs, nutritional supplements, clinical study drugs and others. Identifying other concepts, such as signs and symptoms or diagnoses poses similar difficulties. We refer to the additional terms added to the dictionary as "residual terms" since they represent just a few percent increase in the number of entries.

Semantic lexicon expansion techniques, reviewed comprehensively in [2], try to create an expanded dictionary from a small seed lexicon and a document-set. In general, the techniques strive to discover *patterns* of words that precisely identify elements of a semantic class. These techniques focus on learning the majority of the concepts whereas our approach focuses on discovering the *residual* terms to "complete" the dictionary.

Our algorithm, *SPOT,* differs from known techniques with respect to several factors: it is independent of syntactic analysis (except for whitespace tokenization) and language; it scales to massive amounts of data (e.g. tens of GB, hundreds of millions of progress or telephone notes), it is simple to implement and execute; it is easily customized to a particular scenario under investigation and includes a feedback loop which makes it a natural candidate for an active learning application. An important aspect of our algorithm is that the semantics of a concept is derived from the context in which and by whom it is used. For example, "Starbucks" should be identified as a drug (over the counter brand name for caffeine) when exploring a patient's dizziness, a concept not found in standard dictionaries.

Our methodology resembles the Basilsk algorithm (an example of a weakly supervised bootstrapping method) as described in [2], with key differences in four categories: seeding, scoring, linguistic dependency, and performance. Basilisk starts with a small seed dictionary, (e.g. 10 items) and a corpus of text documents from which nouns (in particular, syntactic roles) are extracted. *SPOT* does not require any syntactic parsing, which enables its application to ill-formed text in any language. In the medical domain this means it does well on "informal English" found in progress notes, not just well formed discharge summaries. It assumes a rather large, well-formed seed dictionary.

SPOT takes advantage of a large initial corpus and large seed dictionary by using a scoring mechanism both for patterns and terms, based on the concepts "*confidence", "support",* and *"term support"*. A *confidence* score indicates how sure one is that a particular pattern is a "good" pattern (one which identifies many known entities as well as potential residual ones). *Support* specifies how often a particular pattern that includes a concept from the seed dictionary appears in the training corpus. *Term support* indicates in how many different patterns a seed term occurred. These three measures allow SPOT to consider hundreds of thousands of patterns and create a ranked list of a few hundred suggested residual entities with high (>75%) precision. Other scoring approaches are also possible, such as mutual information in [3], and are a subject of ongoing research.

Our algorithm was tested on a 3GB corpus of transcribed clinical progress notes from a major health organization, consisting of 1,755,931 documents. The seed dictionary was constructed using the trade names and ingredients as published by the FDA in the Electronic Orange Book (OB) [4]. Different thresholds for support and confidence were examined and a tradeoff in accuracy observed. Results for confidence of 68% and support of 3 will be reported. 25%

Topic: learning algorithm
Preference: oral presentation
Presenter: Anni Coden

of terms in the result set were also contained in a common word dictionary [5] and filtered out before final evaluation, which was done against the seed dictionary and also another dictionary constructed from the proprietary, nonproprietary and substance names as published in DRUGS@FDA [6]. A large percentage of discovered terms – 42% did not belong to either dictionary as such and were manually labeled as shown in Table 1.

Residual terms	3830
Correct	78.30%
Incorrect	21.70%

Table 1: Evaluation of residual terms

In Table 2 we show the reasons why the 78.30% were correct, but could not be identified as such with automatic lookup. In this table, "New valid" refers to real drug names that were spelled accurately, but were not in any dictionary (yet).

Categories	**% of correct terms**	**# of unique terms**
New valid	8.40%	252
Misspellings	76.43%	2292
Elisions	12.60%	378
Abbreviations	2.57%	77

Table 2: Categorization of correct residual terms

SPOT has a strength in identifying misspelled drug names and retrieving residual drugs otherwise unidentified. *SPOT* detects novel drug names previously unknown in either the base or target dictionary – here 252 *new* unique drugs were identified. To try a comparison with a bootstrapping method, which involves syntactic parsing, we applied the Stanford parser [7] on the clinical text. Noun phrase detection was quite poor due to the fact that sentences were not identified correctly because of the ill formed text and domain specific punctuation. Thus entity discovery approaches based on syntactic features will perform very poorly in this domain.

Similar experiments were run on different semantic classes, some rather broad (e.g., diseases and symptoms) and some quite narrow (e.g., family tree labels). For the broad categories scoring is more of an issue, as words are used in many different contexts – e.g., smoker is found to be a member of both the disease and symptom classes. For family tree labels, the tool helped expand our dictionary from 40 terms to over 180 (e.g., mggm = Maternal Great Grand Mother). Since the approach is driven by language snippets commonly used in the context of a particular concept, we made the case that *SPOT* can maintain dictionaries given a very large corpus.

In conclusion, we present our tool SPOT, which uses patterns, scored using confidence, support and term support, to find residual entries missing from dictionaries. It has the advantages of requiring minimal syntactic analysis, yet still performs with high precision.

[1] C. Kuijjer, "Semantic Lexicon Expansion using Bootstrapping and Syntax-based, Contextual Extraction Patterns" at /app.kuijjer.com/static/thesis.pdf/
[2] M. Thelen and E. Riloff, "A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts", Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, 2002
[3] P. Pantel and M. Pennacchiotti. 2008. "Automatically Harvesting and Ontologizing Semantic Relations". In Paul Buitelaar and Philipp Cimiano (Eds.) Ontology Learning and Population: Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text. pp. 171-198. ISBN: 978-1-58603-818-2. IOS Press.
[4] http://www.fda.gov/Drugs/InformationOnDrugs/ucm129689.htm
[5] http://en.wikipedia.org/wiki/Moby_Project
[6] http://www.accessdata.fda.gov/scripts/cder/drugsatfda/
[7] D. Klein and C. D. Manning. 2003. Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

Acknowledgements: We would like to thank Dana Ludwig MD, Diane Oliver MD PhD and Joe Terdiman MD PhD for sharing the data and evaluation of our result.