WORD EMBEDDING AND STATISTICAL BASED METHODS FOR RAPID INDUCTION OF MULTIPLE TAXONOMIES

Lawrence Muchemi [1] and Gregory Grefenstette[2]

[1] University of Nairobi, Kenya & Institut National de Recherche en Informatique et en Automatique INRIA-Saclay, Ile-de-France
[1]lawrence.githiari@inria.fr
[2]Institut National de Recherche en Informatique et en Automatique INRIA-Saclay, Ile-de-France
[2]gregory.grefenstette@inria.fr

ABSTRACT

In this paper we present two methodologies for rapidly inducing multiple subject-specific taxonomies from crawled data. The first method involves a sentence-level words co-occurrence frequency method for building the taxonomy, while the second involves the bootstrapping of a Word2Vec based algorithm with a directed crawler. We exploit the multilingual open-content directory of the World Wide Web, DMOZ[1] to seed the crawl, and the domain name to direct the crawl. This domain corpus is then input to our algorithm that can automatically induce taxonomies. The induced taxonomies provide hierarchical semantic dimensions for the purposes of faceted browsing. As part of an ongoing personal semantics project, we applied the resulting taxonomies to personal social media data (Twitter, Gmail, Facebook, Instagram, Flickr) with an objective of enhancing an individual's exploration of their personal information through faceted searching. We also perform a comprehensive corpus based evaluation of the algorithms based on many datasets drawn from the fields of medicine (diseases) and leisure (hobbies) and show that the induced taxonomies are of high quality

KEYWORDS

Taxonomy, Automatic Taxonomy Induction, Word2vec, Distributional Semantics, Web-crawl, Faceted-search, Personal semantics data

1. INTRODUCTION

Taxonomies are essential for many semantic-based tasks such as content organization, guided-navigation, textual entailment and faceted-search. Taxonomies allow us to refine our searches on shopping and auctions sites, by classifying query results into hierarchic categories, called facets, which can be used to understand and limit the scope of our query. In Enterprise Search systems, facets are the main tools used to find known items. One problem for many ad-hoc or small-scale search applications is that no adequate taxonomies exist because most of the available open source taxonomies are either product search oriented (eg eBay[2], GoogleProducts[3]) or are generic knowledge graphs such as WordNet[4] or Wikipedia knowledge graphs. There is an ever-growing need for simple and robust methodologies for automatic taxonomy construction as for example

[1] https://www.dmoz.org/search?q=knitting and https://www.dmoz.org/search?q=knitting&start=20
[2] http://www.cgmlab.com/ebay-category-tree-donload/
[3] https://support.google.com/merchants/answer/160081?hl=en
[4] http://www.w3.org/2006/03/wn/wn20/download

as evidenced by SemEval-2016 Task-13[5] among others. Spurred by this lack, the field of taxonomy learning has become a prominent branch of taxonomy induction over the last twenty years

The basis of faceted browsing is taxonomies that partition the data using orthogonal or semi-orthogonal semantic facets. The taxonomy facets expose the text's related categories and provide an expanded search. For example in a document retrieval system, a user may request for available documents whose subject is *stitching-styles of cardigans*. If the document space is partitioned by appropriate taxonomies such as *knitting>stitching>stitching-styles and knitting>apparel>cardigan*, the taxonomies will ensure that only documents annotated with category mentions of these taxonomies namely *knitting, stitching, stitching-styles, apparel and cardigan* are retrieved thus limiting the document search space.

For an on-project on indexing and retrieval of personal data, we investigated the availability of such taxonomies for the semantic annotation of personal data obtained from social media applications. We targeted applications such as Twitter, Gmail, Facebook, Instagram, Flickr among others with a view of enhancing document retrieval process with facets from the user point of view on their interests. We were looking for taxonomic descriptions of hobbies and of tasks from everyday life, wishing to apply available open data taxonomies. We found that such taxonomies are generally not available in linked open data sources. For example, out of 267 listed hobbies in the Wikipedia, 121 did not possess category or subcategory listings, so we cannot apply techniques such as converting Wikipedia's graph of categories into a taxonomy, as in MENTA [10]. Further survey on open source taxonomies such as WordNet, eBay, Google-Products, Bing[6] among others show that these taxonomies target products and not personal semantic data. The few that are closely related to personal semantics tasks, such as COELTION[7], which targets classification of Everyday living, are manually developed and therefore not scalable. As a second field of case study we investigated the availability of taxonomies for illnesses. We looked at the Autoimmune Diseases category and could not establish any known or gold standard taxonomies for 157 Autoimmune Diseases.

In general, we were able to confirm that there is an acute shortage of taxonomies that are readily applicable to not only personal semantics data but also to other domains of application and more so for ad-hoc or small-scale search applications. The main challenge therefore that we addressed in this paper is how to rapidly induce taxonomies that structure and classify data. We used both the Autoimmune Diseases and personal semantics applications as our case studies. We therefore embarked on the process of building taxonomies in these two fields where we employed and compared two methods for generation of taxonomies that is sentence-level words co-occurrence frequencies (an extension of the method described in [5]) and Word2Vec based method previously used in the context of lexicography in [12].

In section two, we briefly discuss on related work. We present the main concepts of our two rapid taxonomy induction algorithms in section three. We there after discuss some evaluation experiments and major results in section four. We finish off in section five with conclusions and main contributions of this paper.

2. BACKGROUND AND RELATED WORK

Automatic taxonomy induction from text involves three processes: concept mining, concept relations' discovery, and concept hierarchy building. A comprehensive survey can be found [1] which presents the main approaches to these problems. Statistical and other machine learning based approaches are dominant and they exploit the frequencies of terms and probabilities of co-

[5] http://alt.qcri.org/semeval2016/task13/
[6] http://www.cpcstrategy.com/blog/attachments/taxonomy/
[7] https://coelition.org/business/resources/coel-standard/

occurrence of words within the same window of text. Once the text is obtained in the form of a corpus, various theories such as mutual information, similarity measures, divergence measures, correlation ranking, log-likehood ratio among others, are applied in the concepts-mining, relations-discovery and hierarchy-building stages of automatic taxonomy induction processes.

In machine learning approach, classifiers have extensively been used to discover new relationships based on hand-constructed or automatically discovered textual patterns. For example [2] has presented a probabilistic framework for taxonomy induction in which they exploit the Bayes theorem. The framework defines a set of possible features between pairs of words, for example lexico-syntactic patterns such as those that indicate hypernymy. The framework then seeks evidence from a corpus over other word-pairs with similar features and if a given pair of terms has many occurrences of that feature, then it is concluded that the relationship indicated by the given feature is true. Other researchers such as [3] have introduced methods that combine lexico-syntactic patterns and clustering. The lexico-syntactic patterns include patterns such as *{is-a; such-as; including; especially; called; consists-of}* among others and are obviously language dependent. Clustering then incrementally aggregates terms based on a score indicating semantic distance. In general clustering-based approaches usually represent word contexts as vectors and cluster words based on similarities of the vectors. Through clustering, discoveries of relationships that do not explicitly appear in text are made. Wong [7] has reported a clustering method that relies on agents, known as ANTS that traverse a domain specific corpus to cluster concepts. They use a crawler to build a corpus from which they conduct the clustering process. In general clustering-based approaches face the challenge of appropriately labeling non-leaf clusters thereby amplifying the difficulty of the creation of taxonomies [3]. Further they suffer from a bottleneck of reliance on manual designed and constructed features.

In other approaches, heuristics and statistics have been combined with amazing results. For example [4] reports a heuristic based approach in which they start by extracting domain specific terms from a corpus. They then extract the relationships of the terms from definitions that have been extracted from a corpus, such as Wikipedia, by means of a domain independent classifier. Definitions of the form, A is a/an B form the backbone of the ontology graphs. In the SemEval-2015[8] Task 17 on taxonomy extraction, the winning algorithm by [5] also uses heuristics. The process starts from a given list of terms. By identifying sub-strings inclusion and co-occurrences in Wikipedia sentences, the author generates discrete binary relations of the form A is more general than B and so A is a hypernym of B. In yet another heuristic based work [6], the author uses a combination of techniques and heuristics. These include lexico-syntactic patterns of the part of speech (expressed as regular expressions), morpho-syntactic structure of compound terms where the headword is the more general term of the relationship and a look-up from WordNet.

Our work involved creating many taxonomies for the annotation of personal semantics text data and also illnesses data related to autoimmune diseases with fine-grained facets. Some hobbies such as poi (swinging tethered weights through a variety of rhythmical and geometric patterns) and juskei (throwing a peg over a fixed distance at a stake driven into the ground) are rare while others are difficult even for human experts to easily design (eg do-it-yourself). We therefore required to design a language independent and robust approach that rapidly produces high quality taxonomies.

In this work we adopted two approaches namely, heuristic-based approach (co-occurrence frequencies and substring inclusion) which is an extension of [5] and an extension of word-embedding using word2vec that we explain later. Because the algorithm described in [5] produces discrete binary relations only, we extended on this heuristic and were able to generate

[8] http://alt.qcri.org/semeval2015/task17/

complete taxonomies in the form of directed acyclic graphs. For the word2vec, we used this algorithm to extract domain specific words and phrases in a lexicographic task [12] from which we show how to build a taxonomy hierarchy. Both of these algorithms required vast domain specific corpora and we were able to demonstrate how this is achieved using an open source directory, DMOZ to seed a directed crawler. In order to rapidly induce many taxonomies, we piggybacked a directed web crawler on the taxonomy induction algorithm.

3. METHODOLOGY

In both taxonomy induction methods, we commenced by compiling domain specific text corpora for each of the 157 autoimmune diseases and 266 hobbies making a total of 423 corpora. Building a domain-specific corpora can be achieved by seeding a crawler with urls related to that domain and by providing filters that ensure only web pages of interest are retrieved. However a challenge is encountered in that harvesting these seed urls from the www manually is a very laborious task for multiple domains. We therefore devised a method that provides a linkage between our crawler and an open sourced directory of subject specific links, DMOZ. This therefore provides the urls required for directing the crawling to compose a domain corpus.

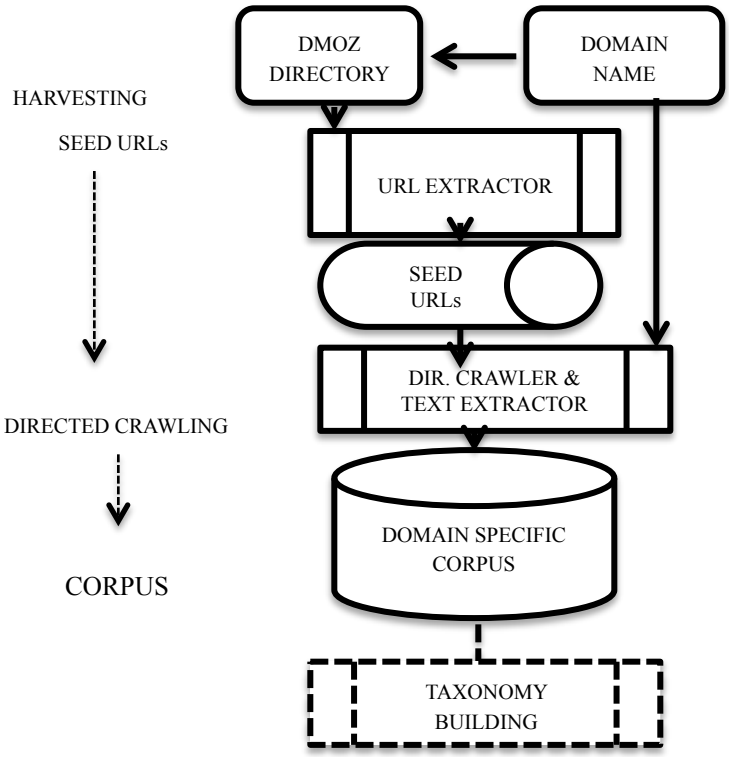


Fig. 1. Rapid Building of Domain Specific Corpora for Multiple Domains

To start the process, we begin with a key domain word. In our experiments, we used the names of hobbies and autoimmune diseases given by the Wikipedia page List of Hobbies and Autoimmune diseases respectively. The word becomes the input in a programmed request to the Open Directory of DMOZ. The request brings back 40 URLs indexed by that word. These URLs become the seed URLs for the directed crawl.

The directed crawl works by picking an uncrawled URL from the set of URLs to crawl initially build from the 40 urls for each domain. A text version of the web page is created using the Unix

lynx command in dump mode. The text is split into sentences[9], and outgoing links are collected. If the text passes a domain filter, then the text is added to the domain corpus and the outgoing links are added to the list of URLs to crawl. Our domain filter is currently set to the initial word used to start the process. More complicated strategies are possible, for example, pre-defining words or patterns specific to the domain [8]. We opted for a conservative approach that works well for specific words such as *Fibromyalgia* or *Gunsmithing* but less well for words, which also have a general meaning such *as Acting*. The crawl stops when a pre-set number of documents, N is added to the domain corpus (we used N=1000), or when the list of URLs left to crawl is empty. To encourage diversity, we also imposed a limitation of 100 documents from the same URL domain (such as *amazon.com*). We then proceeded to apply each of the two methods of taxonomy induction.

3.1 Sentence-level Words Co-occurrence Frequency Method and Subsequences

We now describe the important heuristics and steps necessary in the realization of the fully automatic domain-specific taxonomy generation algorithm. From the onset we defined a 'word' as any stemmed non-stop word, a 'phrase' as any sequence of words between stop-words and a 'term' as any stemmed word or a phrase. Our algorithm relies on two main heuristics and a filter that ensure high quality taxonomies.

The first heuristic is founded on the observation that if two phrases appear in the same sentence, the two phrases are semantically connected. In a number of experiments reported in [5], a term located within a sentence is found to be either more 'general' or more 'specific' compared to another term within the same sentence. In order to find computationally which term is more general than the other, a number of heuristics were tried and the one that seemed to hold true in most texts is the one that if a domain term B co-occurs in the same sentence as a domain term A, B is more likely to be term A's hypernym so long as it appears in more documents than term A.

The other heuristic that was applied to this work is that of subsequences. A subsequence is a sequence contained in or forming part of another sequence. For example, in the sentence

*'**Underwater swimming** on the back has the additional problem of **water** entering the nose.'*

the following relations of the type ***X<broader>Y*** are observed,

-through subsequence : ***underwater >swimming***

- through phrase cooccurrence : ***underwater swimming>water***

The swimming domain specific terms are 'underwater swimming' and 'water'. The terms 'back' and 'nose' though very relevant in this sentence belong to the 'human anatomy' domain and are more salient in that domain. We require further heuristics to separate these domain specific terms and assist in obtaining cleaner taxonomies. After experimentation we obtained a 'terms document-frequency based heuristic' that we explain a little later.

The steps necessary to achieve the automatic domain-specific taxonomy generation are illustrated in the framework found in figure 2.

[9] http://listserv.linguistlist.org/pipermail/corpora/2007-October/010593.html

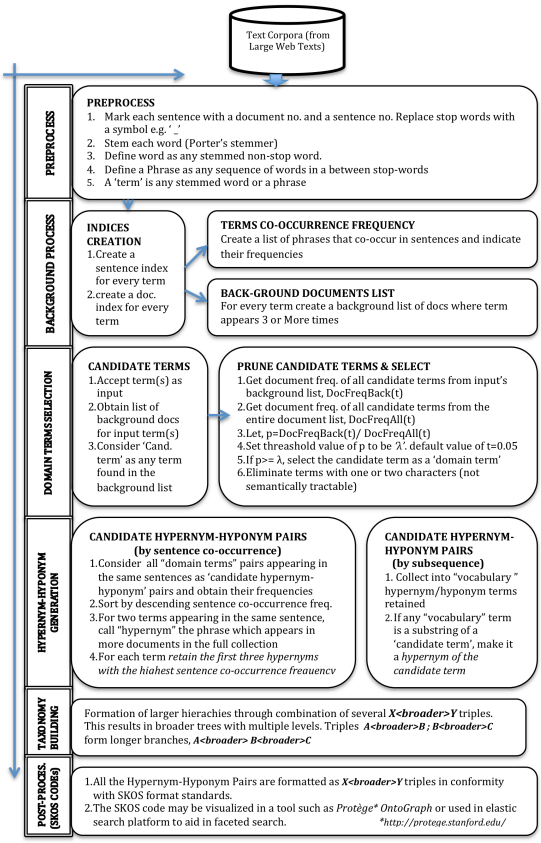


Fig. 2. Sentence Level Phrase Co-occurrence Taxonomy Generation Framework

The process starts by crawling in the web and scrapping large text corpus from the relevant pages as explained earlier. The first phase involves the pre-processing of the text corpora by converting the mined text into 'one sentence per line' corpus and each line marked by document and sentence number tags. The stopwords are then removed and the words stemmed through Porter's stemmer. The un-stemmed form of each word is also retained for the purposes of building a full-words taxonomy as opposed to a stemmed version.

A background processing phase follows the pre-processing one. In this stage each term, a sentence and a document index are created. Further a list of all phrases that co-occur in sentences is created and their frequencies of occurrence indicated. For every term a background list of documents is created. For a given term a document qualifies into this list if contains the term at least three times or more.

The third phase involves harvesting of domain terms. From an initial one (or more) domain specific word supplied by a user, a list of background documents is created by obtaining all the documents where the term(s) appears three or more times. All the terms contained in these background documents are considered 'candidate domain terms'. This is followed by a filtering process of the terms so that we obtain the true domain specific terms. This is done through a 'terms document-frequency based heuristic that applies a threshold, λ to a term's ratio of the document frequency within the background documents dived by the term's frequency in the entire corpus, p. A default value of 0.05 was used in our experiments. Short words of one or two word lengths were also filtered out because in most cases they are semantically intractable.

The fourth phase involves the generation of hypernym-hyponym pairs and determination of which of this is the hypernym. The end result of this phase is a triple of the form 'hypernym-

relation-hyponym' or simply, X<broader>Y triple. Two heuristics are involved in this phase. These are the terms' sentence level terms co-occurrence frequency and terms subsequence relations, which were explained at the beginning of this section (see section 3.1).

The fifth phase involves the formation of larger hierachies through combination of several X<broader>Y triples. This results in broader trees with multiple levels. For example, suppose we had the following triples ***A's<broader>B ; B's<broader>C ; D's<broader>B*** the tree indicated in figure 3 would result.

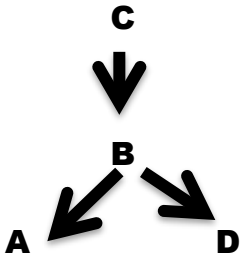


Fig. 3. Taxonomy with Broader and Longer Branches

Finaly an optional post processing that involves conversion in SKOS format and visualization may be done. Through these simple heuristics large taxonomies with high level of precision and recall are achieved.

3.2 Word Embedding in Taxonomy Generation

Two often-used word-embedding methods are Continuous Bag of words (C-BoW) and Skip Gram models introduced in [13] and [14] respectively. The idea behind C-BoW is the utilization of a layered neural network to predict a centre word given some context words while the Skip-Gram model typically takes in one word and tries to predict the closest surrounding words. In both models the words are encoded into real valued vectors of a fixed size for a particular task. The typical dimensions for these vectors range between 50 and 1000 with a width of size 1. The vector values typically represent latent features that are learned by the neural network. It therefore means that words with similar meaning or features will have vectors that are close to each other. To calculate the distance between these vectors, the cosine distance is normally computed.

In our work we used the Skip Gram word2vec word-embedding model to identify terms that are specific to a domain. We utilized the skip gram model where we implemented the word2vec[10] code available in Google code archive. This typically gave use the 50 closest words to the domain name, say the '*Vitiligo*' autoimmune disease. We picked the 25 closest words to the domain name. We found out that the method gives fairly accurate predictions so long as the texts from which the neural network is trained on comes from a narrow domain. This avoids problems of polysemy and synonymy. The details of this domain-specific lexicon identification process and evaluation are found in [12].

Once the lexicon and phrases for a given domain are obtained, we determined the relative frequency of terms within the domain corpus and within a corpus made from a combination of all Wikipedia articles. We named these the technical and background corpus respectively. We considered only the most frequently co-occurring words and phrases (terms). We tabulated the number of co-occurrences for candidate terms, their relative frequencies in the domain (technical) and background corpus along with the respective terms. We build a hierarchy based on the

[10] https://code.google.com/archive/p/word2vec/

principle that more general terms have a higher relative frequency than specific words, hence the more frequent term is a hypernym of the less frequent one. In order to capture more relevant phrases we extract all the terms appearing in the taxonomy build in the first pass and grab any longer phrases that share this vocabulary, so long as they were not captured in the first pass. We obtain their hypernyms (or hyponym) and add it to the taxonomy. The taxonomy build so far is made up of stemmed words. These are converted back to the un-stemmed form to obtain the final version of our taxonomy. These steps are summarized in figure 4 below.

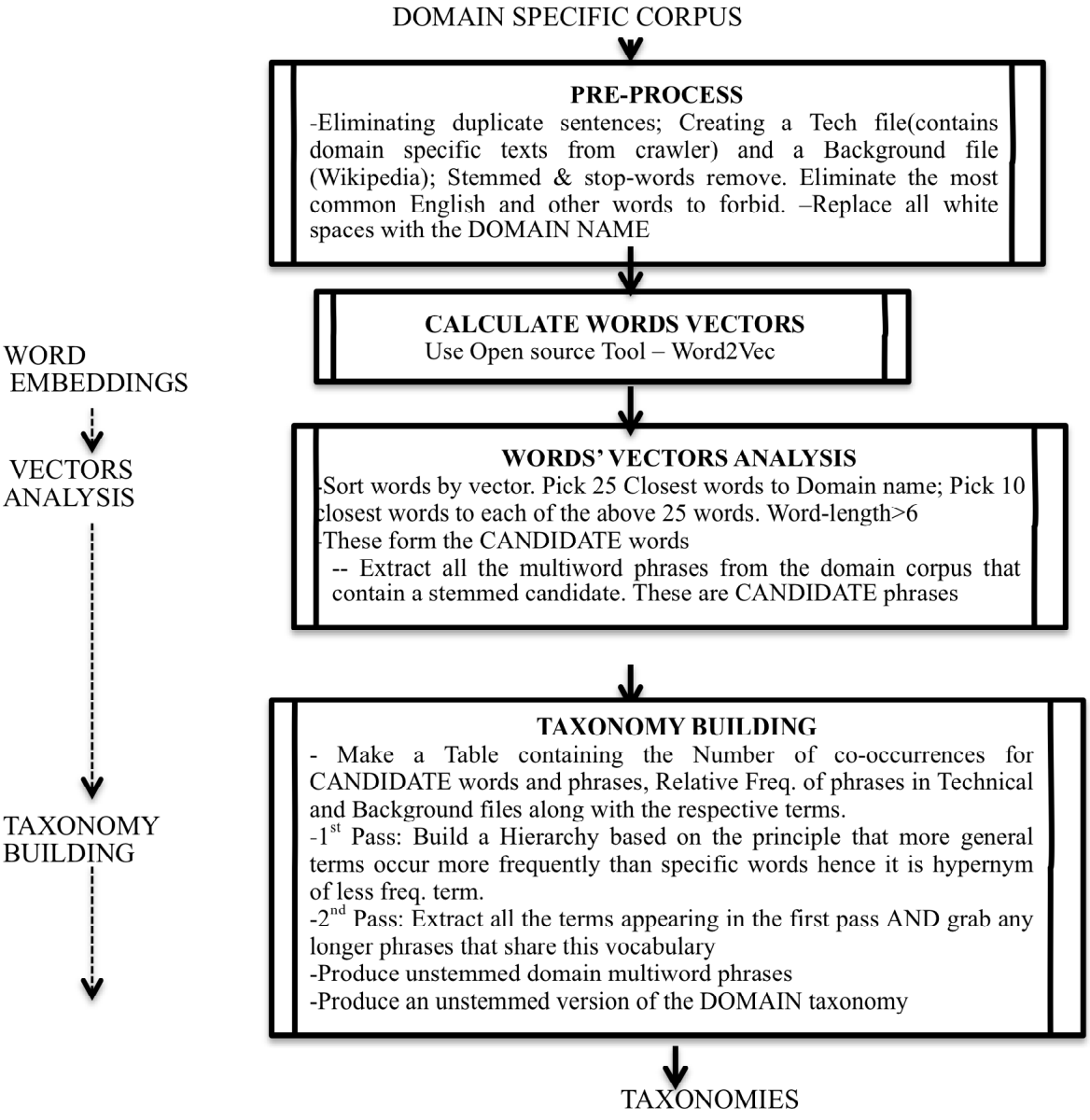


Fig. 4. Rapid Building of Domain Specific Corpora for Multiple Domains

4. EVALUATION

The key objective of our evaluation experiments was to determine the efficacy of the induced taxonomies. Many techniques for evaluating taxonomies exist but among the key ones include:

- Manual evaluation, where experts assess the taxonomies
- Comparison to a gold standard taxonomy or taxonomies generated by baseline algorithms,

- Letting the taxonomies run in an test environment and users give feedback via questionnaires and,
- Evaluation against a corpus such as a document collection

Each of these methods may have some variants in terms of the actual parameters used however, the ultimate objective is to assign some quantitative or qualitative value to the performance and then make comparisons to the state-of-the-art taxonomies.

In our research the goal was to rapidly produce taxonomies for various semantics themes such as hobbies (leisure) and autoimmune diseases (illnesses) and then perform experiments to determine how suitable these taxonomies are to the task of document retrieval. Our ultimate goal is to assist users in browsing and retrieving documents guided by the induced taxonomies.

We targeted domains of interest that are hard to manually evaluate due to scarcity of experts (eg for rare hobbies) or do not have existing gold standards. This then narrowed down our choice of evaluation method to either using the taxonomy in an application environment and assessing its performance through user feedback or evaluating against a corpus derived from independent crowd sourced data. In this paper we present the results from evaluation against many independent crowd sourced corpora. In order to maintain objectivity, we developed our testing corpora from *Reddit*[11] comments, which are crowd sourced on specific themes.

4.1 Experiments

The evaluation task involved the creation of the domain specific taxonomies and evaluating each of these against a text corpus that contains different documents that are known to contain positive example and negative examples. We used the procedures described in section 3 and produced 157 autoimmune diseases taxonomies and 266 hobbies taxonomies making a total of 423 taxonomies in total.

We then gathered *Reddit* comments for a representative sample of 40 taxonomies for the hobbies and 22. We restricted the number of comments to a maximum of 800 per hobby and 300 per an autoimmune illness. These became the positive corpus.

We also generated a negative corpus for every hobby by generating *Reddit* comments that are not related to that hobby. We restricted this to about 3000 documents for each hobby and 2000 documents for each illness. These became the negative corpora for each of the hobbies and illnesses.

The testing procedure consisted of annotating documents from the both positive and negative corpus with facets from the induced taxonomies and recording the true and false positives, and true and false negatives. We defined true positive (TP), false positive (FP), false negative (FN) and true negative(TN) as follows.

TP = Number of documents that were annotated and were supposed to be annotated,
FP = Number of documents that were annotated but were not supposed to be annotated,
FN = Number of documents that were not annotated and should have been annotated,
TN = Number of documents that were not annotated and should not have been annotated

A document was considered annotated if it had at least one matching word with the taxonomy under test.

We then determined Precision, Recall and F1 scores using the general formulae:

[11] https://www.reddit.com/

$$P = TP/(TP+FP)$$
$$R = TP/(TP+FN)$$
$$F_\beta = (1+\beta 2).P.R/((\beta 2.P)+R).$$

To provide a comparison, the tests were repeated but with taxonomies generated from Wikipedia articles and categories where this was available. The results are found in the next section.

4.2 Results for Hobby Activities

Table 1 shows the average performance across the six major hobby categories that we tested. The taxonomies were generated using the *sentence-level words co-occurrence frequency method.* Three hobbies were sampled per category and the results are tabulated here below.

Table 1. Average Performance across the Six major Hobby Categories

Category	*Sample Taxonomies*	*No of Lines*	*Recall*	*Precision*	*F-1*
Games	Boad-games	684	0.848	0.665	0.746
	Racquetball	1905	0.686	0.481	0.566
	Swimming	566	0.848	0.856	0.852
Workmanship	CandleMaking	1213	0.875	0.923	0.899
	LeatherCraft	716	0.613	0.669	0.64
	Amateur Radio	385	0.673	0.869	0.758
Drama & Arts	Dancing	2552	0.85	0.329	0.474
	Calligraphy	7109	0.418	0.471	0.443
	Digital-Arts	282	0.442	0.411	0.426
Clothing & Costumes	Knitting	3101	0.894	0.815	0.852
	Cosplaying	14950	0.690	0.620	0.653
	Crocheting	12155	0.727	0.477	0.576
Knowledge & Creativity	Language Learning	1843	0.812	0.495	0.615
	Cryptography	1830	0.794	0.717	0.754
	Creative Writing	623	0.393	0.717	0.508
Cooking & Brewing	Cooking	4155	0.617	0.567	0.591
	Home Brewing	4258	0.902	0.530	0.667
	Roasting Coffee	1677	0.860	0.561	0.678
	Average	-	**0.719**	**0.621**	**0.685**

The sampled taxonomies fall broadly under 6 major categories namely *Games, Workmanship, Drama & arts, Clothing & costumes, Cooking & Brewing* and *Knowledge & Creativity*. Here we present results for 18 taxonomies. The selected taxonomies included hard-to-generate and rare-hobbies taxonomies on one end and hobbies with elaborate taxonomy facets and therefore easy to generate from human point of view.

These results are further analysed and plotted in figure 5.

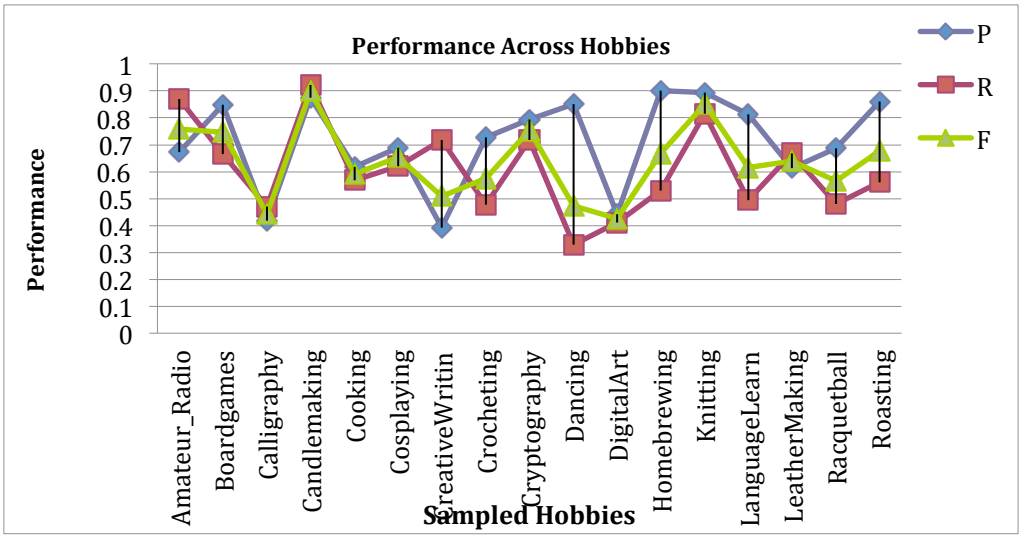


Fig. 5. Performance Across Hobbies Domains

The graph shown in figure 5 shows the performance across various hobbies. The results indicate a consistently high performing algorithm but with several notable exceptions especially in abstract subjects such as arts where the performance is low. A probable reason would be that most texts for the very abstract subjects (from which we developed the corpus) also tend to use very general terms that are also common in English.

4.3 Results for Autoimmune Illnesses

Table 2 presents results for the Autoimmune Illnessess. The taxonomies were generated using the *Word2Vec based algorithm.*

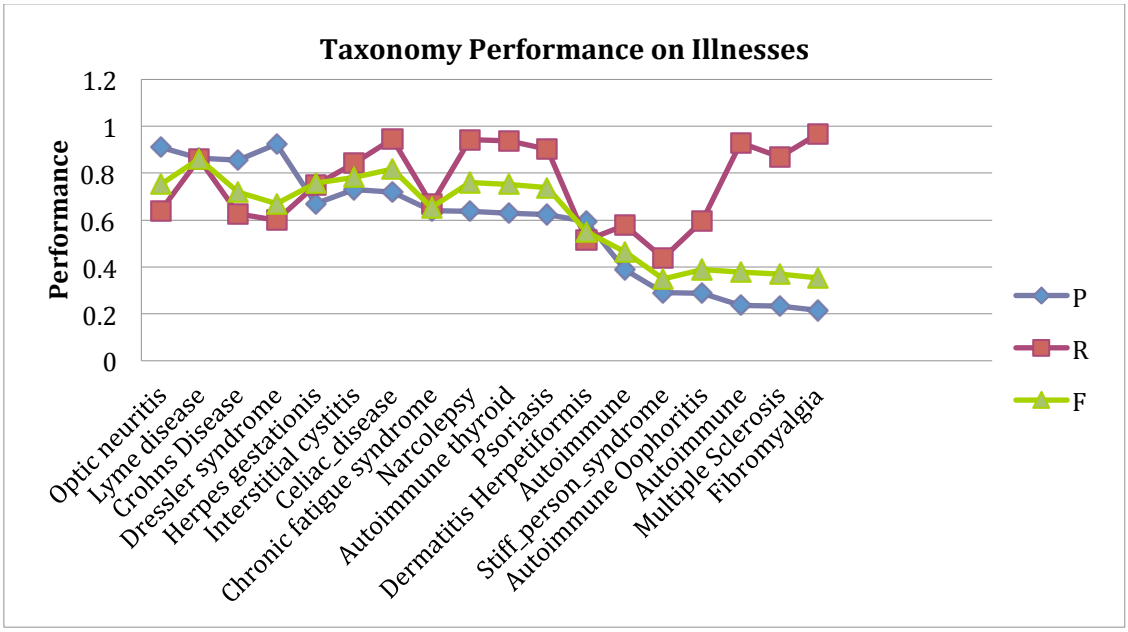


Fig. 6. Performance Across Autoimmune Illnesses Domains

The results from the 157 diseases would be too much to be contained in this paper however, we randomly sampled a total 18 illnesses. One criteria used was that it should have over three hundred crowd sourced comments from the Reddit forum. The detailed results are shown in table 2.

Table 2. Average Performance Across Autoimmune Illnesses Domains

Sample Taxonomies	***No of Lines***	***Recall***	***Precision***	***F-1***
Optic neuritis	1729	0.640	0.911	0.752
Lyme disease	2917	0.860	0.864	0.862
Crohns Disease	887	0.625	0.856	0.720
Dressler syndrome	238	0.598	0.923	0.671
Herpes gestationis	909	0.750	0.669	0.758
Interstitial cystitis	1381	0.841	0.730	0.781
Celiac_disease	1508	0.945	0.719	0.816
Chronic fatigue syndrome	1238	0.668	0.640	0.654
Narcolepsy	2192	0.942	0.638	0.761
Autoimmune thyroid disease	3029	0.938	0.628	0.752
Psoriasis	1844	0.902	0.624	0.738
Dermatitis Herpetiformis	1098	0.513	0.594	0.550
Autoimmune immunodeficiency	872	0.578	0.390	0.465
Stiff person syndrome	973	0.438	0.290	0.349
Autoimmune Oophoritis	956	0.596	0.289	0.389
Autoimmune Dysautonomia	292	0.928	0.237	0.378
Multiple Sclerosis	2160	0.870	0.234	0.369
Fibromyalgia	6253	0.968	0.215	0.352
Average	**1693**	**0.755**	**0.577**	**0.618**

4.4 Comparison with Wikipedia-derived Taxonomies

Table 2 shows a comparison of the performance of some publicly available hobby taxonomies in comparison with some of our taxonomies. We generated linear taxonomies from the Wikipedia acyclic graphs as in MENTA [10] and tested them against the test corpora.

Table 3. Results from Representative Taxonomies

		P	**R**	**F-1**	**Observations**
Knitting	ATC	**0.894**	**0.815**	**0.852**	ATC has higher R
	Wikipedia	0.894	0.648	0.751	
Caving	ATC	0.962	**0.775**	**0.858**	Equal F-score and almost similar F, R
	Wiki	**0.976**	0.766	**0.858**	
Hunting	ATC	**0.983**	0.458	**0.624**	ATC has higher precision and higher F1
	Wiki	0.665	**0.559**	0.607	
Swimming	ATC	**0.848**	**0.856**	**0.852**	ATC has higher precision and higher F1
	Wiki	0.766	0.835	0.799	
Average	ATC	**0.922**	**0.726**	**0.797**	ATC has higher P and R for the compared Taxonomies
	Wiki	0.825	0.702	0.754	

In this evaluation task, we were keen to assess the efficacy of the taxonomies as opposed to an absolute value that would be obtained from a gold-standard. These gold-standard taxonomies are rare and only a baseline taxonomy such as that generated from Wikipedia categories can be used to qualitatively assess the suitability of the usage. We have mainly used a corpus method to assess the usability of their usage. We plan to release the full set of the results with the publication of this paper.

4.5 Four Examples of the Taxonomies Induced (out of the 423)

An Extraction from the Knitting Taxonomy (Porter Stemmed Concepts)

knit>cast-on>sweater>button band
knit>cast-on>sweater>classic irish knit dog sweater
knit>cast-on>sweater>comment question thank
knit>cast-on>sweater>doneg
knit>cast-on>sweater>finish object
knit>cast-on>sweater>finish sweater
knit>cast-on>sweater>knit babi sweater
knit>cast-on>sweater>knit raglan sweater
knit>cast-on>sweater>knit soap sweater
knit>cast-on>sweater>knit sweater
knit>circular knit
knit>circular needl
knit>circular needl>pattern
knit>circular needl>pattern>alissa
knit>circular needl>pattern>beauti yarn
knit>circular needl>pattern>cabl pattern
knit>circular needl>pattern>cardigan knit pattern
knit>circular needl>pattern>chunki arm knit blanket pattern

An Extraction from the Cooking Taxonomy (Porter Stemmed Concepts)

add cup>cook pasta
add salt>half cook
airtight contain>cooki store
allrecip>allrecipes.com
allrecip>allrecip staff
allrecip>cook tip
allrecip>recip box
allrecip staff>cooki dough
allrecip staff>cook question
allrecip staff>halloween cooki
allrecip staff>sugar cooki
avocado>closet cook
avocado>creami avocado
bake cooki>bake cooki set
bake cooki>freshli bake cooki
balsam vinaigrett>kevin lynch said... anonym
balsam vinaigrett>quinoa salad
biryani>electr rice cooker
biryani>pot meal

An Extraction from the Vitiligo Illness Taxonomy (No Porter Stemming)

vitiligo>cure vitiligo
vitiligo>cure vitiligo>cure vitiligo naturally
vitiligo>vitiligo naturally>cure vitiligo naturally
vitiligo>cure vitiligo>cure vitiligo oil
vitiligo>vitiligo patches>cure vitiligo oil
vitiligo>vitiligo photo>cure vitiligo oil
vitiligo>cure vitiligo>curing vitiligo naturally gray hair cure
vitiligo>dark skin
vitiligo>darkening
vitiligo>darker skin types
vitiligo>darker-skinned people
vitiligo>darker skin>darker skin types show maximal responses

An Extraction from the Fibromyalgia Illness Taxonomy (No Porter Stemming)

fibromyalgia>pain>muscular pain>abdominal cramping
fibromyalgia>symptoms>cramps>abdominal cramping
fibromyalgia>symptoms>joint pain>abdominal cramping
fibromyalgia>symptoms>muscular pain>abdominal cramping
fibromyalgia>pain>cramps>abdominal cramping
fibromyalgia>pain>joint pain>abdominal cramping
abnormal sleep affects,fibromyalgia>symptoms>sleep>abnormal sleep affects
abnormal sleep pattern,fibromyalgia>abnormal sleep pattern
fibromyalgia>abnormal sleep pattern>abnormal sleep pattern involving stages
abnormal thinking,fibromyalgia>depression>abnormal thinking
abnormal thinking,fibromyalgia>fatigue>abnormal thinking

As it can be observed from these four samples, the taxonomies are fairly linear and straightforward. These are then converted to SKOS format via simple scripts and incorporated in systems that use RDF data.

5.0 CONCLUSION

We have presented two methodologies for rapidly inducing taxonomies. We have elaborated on how the initial seed words emanating from the Wikipedia list of hobbies are sent to a program that interrogates the Open Directory of DMOZ and obtains the relevant URLs that become the seed to the directed crawler. This is a completely automatic process, whose output is a domain-specific corpus. It is from these corpora that we build our taxonomies. In the first method we showed how to extract the domain terminology by taking advantage of two sets of statistics-based heuristics namely the application of sentence level co-occurrence frequency and term subsequence statistics. In the second method, we have shown how to extract domain terminology using word-embedding vectors. We applied word2vec algorithm to domain-specific corpus and extended this to be able to create hierarchies by using the principle that more general terms have a higher relative frequency than specific words, hence the more frequent term becomes the hypernym of the less frequent one. The frequencies used are the relative frequencies of a domain corpus and background corpus. We also described how we evaluated through a corpus-based method to assess the efficacy of each taxonomy. We further compared our results to some baseline taxonomies generated from Wikipedia categories.

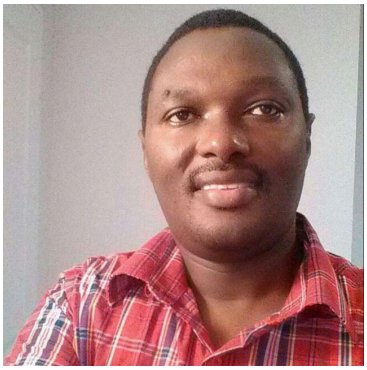
The main contributions of this paper include presenting a completely automated method of building a domain specific corpus through directed crawling where we demonstrated the use of an open source urls directory. We further contribute through the extension of statistical based method that exploits relative co-occurrence frequency and term subsequence statistics through which taxonomies are rapidly induced. Another major contribution was in the use of word2vec algorithm to create lexicon and phrases that form the backbone of taxonomies. We have shown that taxonomies created by both methods are of high quality and can generally support semantic annotation of documents, and subsequent faceted browsing of the annotated content.

References

1. Wong, W., Wei, L., & Bennamoun, M. : Ontology Learning from Text: A Look Back and Into the Future. *ACM Computing Surveys , 44* (20). (2012).
2. Snow, R., Jurafsky, D., & Ng, A. (2006). Semantic Taxonomy Induction from Heterogenous Evidence. *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*
3. Yang, H., & Callan, J. (2009). A metric-based framework for automatic taxonomy induction. *ACL '09 Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, *Vol. 1-1*, pp. 271-279.
4. Navigli, R., Velardi, P., & Faralli, S. (2011). A Graph-Based Algorithm for Inducing Lexical Taxonomies from Scratch. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence* (pp. 1872-1877). Barcelona, Spain: Toby Walsh.
5. Grefenstette, G. (2015). Simple Hypernym Extraction Methods. *HAL-INRIASAC*. Palaiseau, France.
6. Lefever, E. (2015). LT3: A Multi-modular Approach to Automatic Taxonomy Construction. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)-ACL*, (pp. 943–947). Denver, USA.
7. Wong, W. (2009). *Learning Lightweight Ontologies from Text across Different Domains using the Web as Background Knowledge.* Doctoral Thesis, The University of Western Australia, Perth.
8. Bel, N., Papavasiliou, V., Prokopidis, P., Toral, A., Arranz, V.: Mining and exploiting domainspecific corpora in the PANACEA platform. In: The 5th Workshop on Building and Using Comparable Corpora (2012)
9. Wong, W., Liu, W., & Bennamoun, M. (2007, December). Determining termhood for learning domain ontologies using domain prevalence and tendency. In *Proceedings of the sixth Australasian conference on Data mining and analytics-Volume 70* (pp. 47-54). Australian Computer Society, Inc.

10. de Melo, G. and Weikum, G., 2010, October. MENTA: Inducing multilingual taxonomies from Wikipedia. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 1099-1108). ACM.
11. Bordea, G., Buitelaar, P., Faralli, S., & Navigli, R. (2015). SemEval-2015 Task 17: Taxonomy Extraction Evaluation (TExEval). *The 9th International Workshop on Semantic Evaluation*, (p. 902_910). Denver, Colorado, USA.
12. Greffenstte, G., & Muchemi, L. (2016). Determining the Characteristic Vocabulary for a Specialized Dictionary using Word2vec and a Directed Crawler. *International Conference on Language Resources and Evaluation (LREC 2016)-GLOBALEX 2016* (pp. 81-85). Portorož -Slovenia: Kernerman, Ilan et al.
13. Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality". In NIPS, pp. 3111-9. 2013
14. Pennington, Jeffrey, Richard Socher, and Christopher D. Manning "GloVe: Global Vectors for Word Representation". In: Proceedings of EMNLP 2014. pp. 1532–1543. 2014.

Gregory Grefenstette Is an Advanced Researcher at INRIA Saclay, Ile-de-France within the TAO team. He leads the Personal Semantics project. An expert in information retrieval and natural language processing, Grefenstette established the field of Cross Language Information Retrieval by creating its first Workshop at SIGIR'96. He is also one of the pioneers of distributional semantics, following his PhD work « Exploring Automatic Thesaurus Generation » (Kluwer, 1994)
Between 2008 and 2013 he was a Scientific Director at Exalead and a in the period 2004 and 2008, he as a Senior research scientist at the CEA LIST Alternative Energies and Atomic Energy Commission (French: Commissariat à l'énergie atomique et aux énergies alternatives)



Lawrence Muchemi holds an MPhil (Eng) and PhD in Computer Science and lectures at the University of Nairobi, Kenya. He is currently a Post Doctoral researcher at the Institut National de Recherche en Informatique et en Automatique (INRIA Saclay), Ile-de-France within the Personal Semantics project of the TAO team. He specializes in Machine learning technics and more so for Natural language Processing. He has authored many academic papers available here and also published the book, Natural Language Access to Database: Ontology Concept Mapping Approach available in Amazon and Lambert Academic Publishers.
He is an experienced Artificial Intelligence (AI) Expert having started as an Engineer in 1995. He has taught at various universities in Kenya, which include JKUAT, (1999-2000) Africa Nazarene University (2000-2006) where he was the head of the department and currently at the University of Nairobi. (2006-Present)