A word-sense representation for documents based on information value and word sense disambiguation

Leonardo Lazzaro [*], **Julián Peller** [*] , **Juan Manuel Pérez** [*]

[*]Universidad de Buenos Aires, Buenos Aires, Argentina

In this project we analyzed some Lenin works in order to identify the presence of some topics of his biography. To achieve this, we defined and implemented a simple metric distance between a word-sense and a document and confronted this metric using a evaluation set of word-senses against a reduced set of documents with well known topics. During this work, we generated an indexed database of Lenin's complete works from a web site[1] and we applied a series of syntactic, semantic and lexicographic tools in order to model the similarity between a document and a word-sense. We applied information value[2] and a simple algorithm of word sense disambiguation[6] based on Wordnet[5] in order to generate the *top senses* (or principal senses) of a document. Finally, we compare them against some semantic senses of interest (for example: revolution, war, philosophy, etc).

information value | word sense disambiguation | natural language processing | information retrieval | polysemy

Introduction

In this article we propose a simple distance metric between a word-sense and a text document using different well known tools from natural language processing. To achieve this, we extract a hierarchy of the most representative words of a document (*top words*) using the concept of *information value*, a statistical linguistic model based on Shannon's information theory and proposed by Zanette et al. in[2]. Briefly, information value allows us to represent a document like a ranked set of words.

But this weighted set of top words has the problem of polysemy, typical in the field of nlp. Polysemy occurs when a term has more than one sense or meaning, for example: the term 'bank' can refer to a financial institution or to a land formation, thus, it is polysemic. In order to resolve this limitation, we implement a basic algorithm of word-sense disambiguation based on wordnet's synsets and we apply it over the top words, using as disambiguation context the whole document. We obtain, then, the *top senses* of a document, this is, a ranked set of wordnet synsets that work like a representation of the given document.

Finally, we utilize different wordnet predefined metrics of similarity between synsets to define our own similarities between the top-sense representation of a document and any given word-sense (wordnet synset). In order to evaluate the behaviour of this metrics, we test the distances of the top-senses representation of some well known documents with very specific topics against a bounded set of relevant evaluation word-senses.

Method

Text Corpus. We selected some starred works from Lenin from the "Marxists Internet Archive"[1]. Those works are very different in topic, some of them being strictly phylosophical essays (such as "Materialism and Empiriocriticism") and other being more political, such as "State and Revolution", and finally an economic essay called "Imperialism, the highest stage of capitalism".

- What is to be done(1901)
- The Agrarian Programme of Social-Democracy in the First Russian Revolution, 1905-1907 (1907)
- Materialism and Empiriocriticism (1908)
- Imperialism, the highest stage of capitalism (1916)
- State and Revolution (1917)

For this purpose, we developed a web crawler to fetch the works from the [1], but ended up using only these works.

Semantic Analysis.

Information value

The information value of a word for a given text is defined and tested by Zanette *et al.* in [2] and [3]. It is a real value that quantifies how much information a word carries according to the text. Let T be a text, and T' a random shuffled version of T, and w a word appearing in T. By $S_T(w)$ we denote the Shannon Entropy of w with respect to T (See appendix for more info).

The information value of w in T is

$$IV_T(w) = f_w * |S_T(w) - S_{T'}(w)| \quad [1]$$

where f_w is the frequency of appearance of w in T. This value captures, in some way, whether a word has a distribution in text that is not exactly uniform; an order of the word in the text that brings some sense to it.

Ordering the words by their information value results in a list of the text's most representative words. In the context of our research, we call the first n words of a text -with n fixed-, ordered by it's information value, the *top words* of the document.

That's it, if we have all the words sorted as $w_1, w_2, \ldots, w_M$ such that $i \leq j \Rightarrow IV_T(w_i) >= IV_T(w_j)$, then, the n top words are $w_1, w_2, \ldots w_n$

Reserved for Publication Footnotes

Wordnet synsets & similarities

Having the *top-words* representation of a document, we were interested in measuring the distance or similarity of that text to different words. In order to address this problem we used different similarity functions defined over wordnet in the python's library nltk[1]. Wordnet is a lexical resource known as thesaurus, i. e, a data base of terms grouped as sets of synonyms (synsets), each of which have a definition and different semantic relations with others synsets (for example: hypernymy, hyponymy, meronymy, etc). The similarity functions are defined over this network of semantic relations.

This functions aren' t defined over words but over synsets. A synset, briefly, is an abstract notion of sense, meaning or concept that is independent from it's linguistic formulation or spelling. Using synsets instead of words, the wordnet thesaurus solves two problems of ambiguity limited to the linguistic domain: polisemy and synonymy. Polisemy occurs when a unique term may refer to more than one concept (for example, bank may refer to a financial institution or to a land formation. In the other hand, synonymy occurs when different terms refer to the same concept (for example 'car' and 'automobile' refers, both, to the same concept).

The similarities defined in the nltk-api are the following: Path similarity, Leacock-Chodorow similarity, Wu-Palmer similarity, Resnik similarity, Jiang-Conrath similarity and Lin similarity. The former three are analytical, this is, their calculus is based only in the semantic relations of the thesaurus, meanwhile, the last three are analytical but also use statistics from a corpus.[7][8]. In this project, we present the results measured with path similarity because the idea is to present a proof of concept of a representation for documents (and not a 'productive' one). This similarity represents the distance between two nodes in Wordnet based on the shortest path that connects the senses in the is-a (hypernym/hypnoym) taxonomy.

Word sense disambiguation

Word sense disambiguation involves the association of a given word in a text or discourse with a definition or meaning.

Our approach for WSD was a classical Lesk[6] implementation that uses wordnet as a definition source with some tweaks.
Basically Lesk is based on the idea that words that co–occur in a sentence are being used to refer to the same topic.
The Lesk algorithm takes the word definition or gloss that we want to disambiguate and it compares it gloss against the glosses of the other words in the context, then a word is assigned that sense whose gloss shares the largest number of words in common with the glosses of the other words.
This algorithm suffers the problem that relies on the word definition and definitions usually uses as less words as possible. Not only that sometimes two different words, but related ones, uses no words in common in their definitions.
As we use the nltk wordnet we immediately notice an improvement using the rich set of relationships that Wordnet gives to us, so we added to our Lesk implementation word stems, hypernyms, hyponyms. For the problem of short word description we modified the algorithm to use the definition of the word to disambiguate and to compare it against the context words directly. Finally our algorithm chooses the best synset from wordnet based on the context words.
We had choosen the whole text as context for all runs.

Top-sense definition

In a first approach, we pretend to measure the distance between a document (represented by it's *top words*) and different words. To address this problem using the wordnet distances between synsets described before, we modeled the notion of 'word' as the sum of all its synsets. The results were pretty demotivating, so we refocused our energy in order to define our problem in the space of word-senses (and no more in the space of words).

Using the WSD algorithm described in the previous section, we defined the notion of *top senses* of a document as the word-senses resulting from the disambiguation of the *top words* of the documents obtained with Zannete's information value approach.

[1] http://www.nltk.org/

Handpicked concepts to compare against top senses

Synset	Gloss
war	the waging of armed conflict against an enemy
strategy	the branch of military science dealing with military command and the planning and conduct of a war
practice	translating an idea into action
money	the most common medium of exchange; functions as legal tender
imperialism	a political orientation that advocates imperial interests
proletarian	a member of the working class (not necessarily employed)
revolution	a drastic and far-reaching change in ways of thinking and behaving
capitalism	an economic system based on private ownership of capital
materialism	(philosophy) the philosophical theory that matter is the only reality
theory	a well-substantiated explanation of some aspect of the natural world
philosophy	the rational investigation of questions about existence and knowledge and ethics
idealism	(philosophy) the philosophical theory that ideas are the only reality
lion	large gregarious predatory f of Africa and India having a tawny coat with a shaggy mane in the male
toilet	the act of dressing and preparing yourself
television_receiver	an electronic device that receives television signals and displays them on a screen
hair	a covering for the body (or parts of it) consisting of a dense growth of threadlike structures
sugar	a white crystal carbohydrate used as a sweetener and preservative
car	a motor vehicle with four wheels; usually propelled by an internal combustion engine

Top sense for all Lenin works

The State and Revolution	Imperialism,the Highest ..	Materialism and ...	What Is To Be Done?	The Agrarian Programme ...
marx	enterprise	physics	local_anesthetic	rent
revolution	depository_financial...	sociable	organization	land
plant	imperialism	space	marriage	peasant
engels	universe	time	criticism	capital
club	million	berkeley	revolutionist	revolution
power	colony	truth	freedom	marx
party	colonial	aim	newspaper	people
sociable	product	sensation	worker	nationalization
experience	penny	experience	consciousness	restoration
bernstein	company	engels	newspaper	right
question	capitalist	motion	businessperson	private
equality	monopoly	marx	mass	farming
phase	area	book	motion	free
commune	policy	plant	german	central
one	power	general	goverment	land
democracy	banks	helmholtz	political_orientation	capitalism
class	worker	neccesity	party	democracy
law	imperialist	law	secret	landlord
communism	company	sensation	working	farming
peer	boards	energy	class	agency

The table shows $top-senses$ obtained from the choosen works. Each $top-sense$ is ordered according to the information value score that we obtained previous to apply the WSD algorithm. Most of the senses found by our algorithm were the senses related to the works, except for some of them. On Whats is to be done? local_anesthetic was matched as sense. On State and revolution Marx was matched with a United States comedian and the same result we got with Berstein that was matched with a composer

Similarities with handpicked concepts

The State and Revolution

Synset	Similarity
revolution	1.0
imperialism	0.333
proletarian	0.25
war	0.166
practice	0.166
money	0.125
philosophy	0.125
materialism	0.125
strategy	0.125
idealism	0.125
capitalism	0.125
toilet	0.111
hair	0.111
television_receiver	0.1
car	0.1
theory	0.1
lion	0.0833
sugar	0.0833

Imperialism, the Highest St...

Synset	Similarity
proletarian	0.25
hair	0.2
money	0.166
practice	0.166
toilet	0.142
war	0.142
revolution	0.125
theory	0.125
capitalism	0.125
television_receiver	0.111
car	0.111
philosophy	0.111
materialism	0.111
strategy	0.111
idealism	0.111
imperialism	0.1
lion	0.0833
sugar	0.0833

Materialism and Empirio...

Synset	Similarity
revolution	0.25
proletarian	0.166
philosophy	0.166
war	0.166
strategy	0.166
practice	0.166
materialism	0.142
idealism	0.142
money	0.125
imperialism	0.125
theory	0.125
hair	0.125
capitalism	0.125
toilet	0.11
television_receiver	0.1
car	0.1
sugar	0.090
lion	0.071

What Is To Be Done?

Synset	Similarity
imperialism.	0.5
proletarian	0.25
practice	0.16
money	0.14
war	0.14
capitalism	0.14
revolution	0.125
hair	0.125
television_receiver	0.11
toilet	0.11
car	0.11
philosophy.	0.11
materialism.	0.11
strategy.	0.11
idealism.	0.11
theory.	0.11
sugar	0.090
lion	0.083

The Agrarian Programme...

Synset	Similarity
revolution	1.0
capitalism	1.0
proletarian	0.25
war	0.2
practice	0.166
hair	0.142
money	0.125
toilet	0.111
philosophy	0.111
materialism	0.111
strategy	0.111
idealism	0.111
imperialism	0.111
television_receiver	0.1
car	0.1
theory	0.1
sugar	0.0909
lion	0.0714

Word-senses evaluation set

Based on this definition, we evaluate the resultant *top senses* of some texts and then check the results of the different similarities with well known concepts in the next sections.

The selected concepts are shown in the table below. We can distinguish three subsets in this selection: concepts related to practical situations like geopolitics and state management (like 'war', 'strategy', 'money'), concepts more related to the theoretical world (like 'philosophy', 'idealism', 'theory') and finally, concepts not related to Lenin's works (like 'sugar', 'car', 'television_receiver'). The motivation behind the selection of the two first groups is to check some easy hipotheses (such as "State and Revolution" talks a lot about "state" and "revolution", the motivation behind the third group is to check our methods looking for false positives, i.e., a high ranking in the metrics for a concept assumed not related to the works. The distinction between the two first groups and the third in pretty taxative, while the distinction between first and second group is more smooth.

Finally, we define the similarity between a document and a word-sense as follows:

$$sim(s,T) = max\left\{path(s,s'), s' \in TS(T)\right\} \qquad \textbf{[3]}$$

Results

This section is divided in two: first, we present the *top senses* of the documents listed in "Text corpus"; second, we present the similarities between the documents and the word-senses evaluation set.

In Top Senses table we can observe that some senses like local_anesthetic was matched on Whats to be done? as first result.

Also in State and revolution Marx was matched with a United States comedian and a similar result we got with Berstein that was matched with a composer, however in the case of Engels it was matched to a socialist who wrote the Communist Manifesto with Karl Marx.

Similarities tables shows that, in most of the works, synsets *lion*, *sugar*, *hair*, *car* and *television_receiver* had a lower similarity in most of the works.

One strange result was *Hair* that matched second in similarity score in Imperialism, the Highest St....

Discussion

Top senses of a text. The method used here, as explained before, was intended just to present the model of creating a list of top senses out of a list of top words. In spite of being quite elemental, the method worked fairly well with most texts, having some expected problems.

The main issues concern about Lesk Algorithm's flaws. One of these is the sensitivity of the context, for which the appearance or not of a single word in the gloss/context can radically change the selected sense for the word. Moreover, we strongly depend on glosses defined by Wordnet, which tend to be short.

Another issue is Wordnet itself. While it has overwhelming information for nouns and their relationships, the resources for verbs and adjectives are not as extensive. This fact was a problem not only for the construction of the glosses for Lesk, but also for measuring distances between synsets and works.

Wordnet's synset distances. As in the WSD problem, we also here just present a simple model to the problem of measuring a distance between a document and a -synset. The path similarity is a very simple similarity measure, which is based only in the distance between two synsets. Other

For "State and Revolution", the distance worked as expected, matching with a high similarity those concepts highly related to the topics of the text, and putting down those unrelated. However, we must note that many synsets were extremely close of "law.n.04", which is a problem carried by a wrong disambiguation (the definition for it has to do with "natural laws", nothing to do with the text).

In "Imperialism ... ", we suffered the appearance of a very general sense (universe.n.01) which made a nonrelated sense as hair to appear second in the ranking. Also, a very characteristic concept of the as **capitalism** appears unrelated, because it is not very close (in the hypernyms/hyponyms taxonomy) to the top senses of this work. "Materialism ... " had the same problem but the other way around: **revolution** was just very close to **experience**, but had nothing to do with the text in general.

Many of these problems come from the fact that path similarity is not a good measure of semantic similarity. Of course, two senses which have a high path similarity are similar (they are very close in terms of hypernyms/hyponyms), but the other way round is not valid. For instance, meronimy is not taken into account in this metric.

Notice that, in order to use the path similarity, we had to select only noun senses for the top words. This is because verbs, adjectives and nouns are in different taxonomies in Wordnet, and hence are not comparable with this similarity.

Conclussion & Future work

The original scope of this investigation was the definition of a measure of distance between the works of a whole year and a concept; so we could measure whether the literary production of Lenin in some year was close to "revolution", "theory", etc. With this purpose, we defined the **top senses** of a text, and measured the distance between them and other senses in a topological way.

Although the final results were not the expected, some tools we developed in our way proved to be promising. The hybrid method for extracting top senses of the text based in entropy analisis plus sense disambiguation worked in an acceptable fashion, with lot of room for improvement by using better WSD algorithms. In the other hand, the topological representation of a text as a set of senses has yet to experimented with better semantic similarity functions than the one used here.

Acknowledgments

Thanks to Marxists Internet Archive for helping us retrieve the Lenin's complete works.

Appendix: Entropy of a word in a text

Let's recall from [2] and [3] the concept of information value of a word. Let T be a text, which has been split into P pieces $T_1, T_2, \dots, T_P$, w be a word in T, and f_i the frequency of appearance of w in T_i.

For each part $1 \leq i \leq P$, we could define:

$$p_i = \frac{f_i}{\Sigma_{j=1}^{P} f_i} \qquad [4]$$

This quantity stands for probability of finding w in T_i, given that it is present in the text. So, we define the Shannon information entropy of w in T as

$$S = -\frac{1}{lnP} \sum_{i=1}^{P} p_i ln p_i \qquad [5]$$

Appendix: Optimal window size selection

As [2] mentions, we should consider the window size which maximizes the information value per word, for some sense of it.

Let's suppose we have a document, and a window size W_S for which the information value algorithm returns a list $\{(w_1, v_1), (w_2, v_2), \ldots, (w_n, v_n)\}$ of word-information value pairs, in decreasing order of information value so that $v_1 \geq v_2 \geq \cdots \geq v_n$. How would we assign some value to this pair of document-window size so we can choose the optimal window size?

One possibility would be to take the greater information value, in this case v_1. This approach, however, gives all the value just to a single word. Another chance would be to take the average of all the information values, but this leads to a lot of noise added because of words without any value.

The chosen approach is a tradeoff between both. We take the average not of all the words, but of a percentage of the most significative words. We call this percentage the *sum threshold*. If n is the number of words in the document, and $0 \leq ST \leq 1$ is the sum threshold, we calculate

$$\frac{\sum_{i=1}^{[STn]} v_i}{n} \qquad [6]$$

The threshold value we used is $ST = 0.001$, so it would consider only the $\frac{n}{1000}$ most significant words. The reason behind this was that it fitted the results from [2] for the texts The origin of the species, Moby Dick, and The Analysis of the mind.

1. Marxists Internet Archive (http://www.marxists.org)
2. M.A. Montemurro & D. H. Zanette, 2009, The statistics of meaning: Darwin, Gibbon and Moby Dick, Significance, Dec. 2009, 165-169.
3. M.A. Montemurro & D. H. Zanette, 2001, Entropic analysis of the role of words in literary texts, Adv. Complex Systems Vol. 5, September 27th 2001
4. M.A. Montemurro & D. H. Zanette, 2010, Towards the quantification of the semantic information encoded in written language, Adv. Complex Systems Vol. 13.
5. George A. Miller, 1995, WordNet: A Lexical Database for English , COMMUNICATIONS OF THE ACM November 1995/Vol. 38, No. 11
6. M. Lesk., 1986, Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone. In Proceedings of SIGDOC '86.
7. Budanitsky & Hirst, 2004, Evaluating WordNet-based Measures of Lexical Semantic Relatedness
8. Budanitsky & Hirst, 2001, Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures