Research Notes

Nicholas Monath, Niklas Shulze, Klim Zaporojets

March 14, 2014

Please be sure to provide links to the sources you take notes from either in the form of a link to the webpage or as a Bibtex citation

Possible Future Points of Research/Papers to Read:

1. *Semantic Measures for the Comparison of Units of Language, Concepts or Instances from Text and Knowledge Representation Analysis* by Harispe, Ranwez, Janaqi, Montmain
2. *Introduction to Information Retrieval* by Christopher Manning
3. *Semantic distance in WordNet: An experimental, application-oriented evaluation of ve measures* by Budanitsky and Hirst (M)(S)
4. *An improved semantic similarity measure for document clustering based on topic maps* (2013) by Muhammad Rafi, Mohammad Shahid Shaikh(S)(Z)
5. *Comparing taxonomies for organising collections of documents* (2012) by Samuel Fernando, Mark Hall, Eneko Agirre, Aitor Soroa, Paul Clough, Mark Stevenson (Z)(M)
6. *Entity Disambiguation with Freebase* (2012) by Zhicheng Zheng, Xiance Si, Fangtao Li, Edward Y. Chang, and Xiaoyan Zhu (M)(S)
7. *Constructing a Focused Taxonomy from aDocument Collection* (2013) by Olena Medelyan, Steve Manion, Jeen Broekstra, Anna Divoli, Anna-Lan Huang, Ian H. Witten (M)(S)
8. *A Taxonomy based Semantic Similarity of Documents using the Cosine Measure* (2009) by Ainura Madylova (S)(Z)
9. *Learning Semantic Similarity* (2002) by Jaz Kandola and others (Z)(M)
10. *Learning the Similarity of Documents: An Information-Geometric Approach to Document Retrieval and Categorization* (2000) by Thomas Joffman (M)(S)

11. *Document Representation and Clustering with WordNet Based Similarity Rough Set Model* (2011) by Nguyen Chi Thanh and Koichi Yamada (S)(Z)

12. *Measuring Semantic Similarity between Words Using Web Documents* (2010) by Sheetal A. Takale and other (Z)(M)

13. *Ranking of Web Documents using Semantic Similarity* (2013) by Poonam Chahal, Manjeet Singh, Suresh Kumar (M)(S)

14. *Unsupervised Semantic Similarity Computation Between Terms Using Web Documents* (2010) by Elias Iosif (S)(Z)

15. *Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web* (2005) by Giannis Varelas and others (Z)(M)

16. *Pairwise Document Similarity in Large Collections with MapReduce* (2008) by T Elsayed and others (M)(S)

17. *Effective Measures for Inter-Document Similarity* (2013) by John S. Whissell, Charles L.A. Clarke (S)(Z)

18. *A New Sufix Tree Similarity Measure for Document Clustering* (2007) by Hung Chim and others (Z)(M)

19. *Exploring the Similarity between Social Knowledge Sources and Twitter for Cross-domain Topic Classification* (2012) by Andrea Varga and others (M)(S)

20. *Similarity Measures for Text Document Clustering* (2008) by Anna Huang (S)(Z)

21. *Algorithmic Detection of Semantic Similarity* (2005) by Ana G. Maguitman and others (Z)(M)

22. *Using aWikipedia-based Semantic Relatedness Measure for Document Clustering* (2011) by M Yazdani and others (S)(Z)

23. *Wordnet-based metrics do not seem to help document clustering* (Between 2009 and 2010) by Alexandre Passos and others (M)(Z)

24. *Link-Based Similarity Measures Using Reachability Vectors* by Seok-Ho Yoon, Ji-Soo Kim, Jiwoon Ha, Sang-Wook Kim, Minsoo Ryu, and Ho-Jin Choi

25. *Latent Dirichlet Allocation* by David M. Blei, Andrew Ng, Michael Jordan

26. *Measures of semantic similarity and relatedness in the biomedical domain* by Ted Pedersen, Serguei V.S. Pakhomov Siddharth Patwardhan Christopher G. Chute

27. *Using semi-structured data for assessing research paper similarity* by Germn Hurtado Martna, Steven Schockaert, Chris Cornelis, Helga Naessens

28. *Finding similar research papers using language models* by Germn Hurtado Martna, Steven Schockaert, Chris Cornelis, Helga Naessens

29. *Integrating Multiple Document Features in Language Models for Expert Finding* by Jianhan Zhu1 Xiangji Huang Dawei Song Stefan Ruger

30. *A Language Modeling Approach to Information Retrieval* by Jay M. Ponte and W. Bruce Croft

1 Semantic Similarity

- **Semantic measures:** mathematical tools used to estimate the strength of the semantic relationship between units of language, concepts or instances, through a numerical description obtained according to the comparison of information formally or implicitly supporting their meaning or describing their nature
- **Semantic similarity**: measures the likeness of terms, words, documents (or any objects which can be characterized through semantics). The likeness of compared objects is based on their meaning or semantic content, as opposed to similarity which can be estimated regarding their syntactical representation (e.g. their string format).
- An **ontology** formally represents knowledge as a set of concepts within a domain, using a shared vocabulary to denote the types, properties and interrelationships of those concepts
- **Semantic similarity** can be estimated for instance by defining a topological similarity, by using **ontologies** to define a distance between terms/concepts
  - A naive metric for the comparison of concepts ordered in a partially ordered set and represented as nodes of a directed acyclic graph (e.g., a taxonomy), would be the minimal distance in terms of edges composing the shortest-path linking the two concept nodes. Based on text analyses, semantic relatedness/distance between units of language (e.g., words, sentences) can also be estimated using statistical means such as a vector space model to correlate words and textual contexts from a suitable text corpus (co-occurrence).
- Note the difference between semantic *similarity* and semantic *antonymy* (how *unrelated* things are) and semantic **meronymy**
  - A **meronym** denotes a constituent part of, or a member of something. For example, "finger" is a meronym of "hand" because a finger is part of a hand. Similarly, "wheels" is a meronym of "automobile".

1.1 Measures

- Two main approaches to measuring the similarity of ontological concepts: **edge-based** and **node-based**
  - Edge-based: which use the edges and their types as the data source
  - Node-based: in which the main data sources are the nodes and their properties.
- Other measures calculate the similarity between *ontological instances*:
  - Pairwise: measure functional similarity between two instances by combining the semantic similarities of the concepts they represent
  - Groupwise: calculate the similarity directly not combining the semantic similarities of the concepts they represent
- There are also a number of statical similarity approaches such as: Latent semantic analysis, Pointwise mutual information, etc. (see article for more information)

2 Information Retrieval by Christopher Manning

- **Information retrieval** (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).
- **Unstructured data**: refers to data which does not have clear, semantically overt, easy-for-a-computer structure[1]

3 Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures

- Compares five measures of **semantic distance**. Note the subtle differences between **semantic distance**, **semantic relatedness**, and **semantic similarity**
  - **S**emantically similar lexical items are connected by a 'virtue of their likeness' (bank & trust & company)
  - **Semantic relatedness**is more general. Lexical items that are semantically similar are also semantically related, but items connected by a relationship like meronymy (car & wheel) or antonymy (theist & agnostic) or a functional relation are also considered related (paper & pencil).
  - **Semantic distance** is the inverse of relatedness
- Five measures of similarity:
  - **Hirst-St-Onge**: Two lexicalized concepts are semantically close if their WordNet synsets are connected by a path that is not too long and that "does not change direction too often". Given by:

    $$rel_{HS}(c_1, c_2) = C - \text{path length} - k \times d$$

    where d is the number of changes of direction in the path and C and k are constants.
  - **Leacock-Chodorow**: Given two lexical entries c_1, c_2, the function $len(c_1, c_2)$ is the shortest path between the synsets of c_1 and c_2 using only IS-A (hyponymity). The semantic similarity is given by:

    $$sim_{LC} = -\log \frac{len(c_1, c_2)}{2D}$$

    where D is the overall depth of taxonomy.
  - **Resnik**: The function $lso(c_1, c_2)$ defines the lowest super-ordinate (most ommon specific subsumed) of the two lexical entries (basically their common ancestor that is deepest in the tree). The semantic similarity is given by:

    $$sim_R(c_1, c_2) = -\log p(lso(c_1, c_2))$$

[1] Note how most text we would want to search is *semistructured*, that is it has tags such as a title or headings etc

where $p(c)$ is the probability of encountering an instance of a sunset c in some specific corpus

- **Jiang-Conrath**: Defines semantic *distance* as:

$$dist_{JC}(c_1, c_2) = 2log(p(lso(c_1, c_2))) - (log(p(c_1)) + log(p(c_2)))$$

- **Lin**'s approach is

$$\frac{2log(p(lso(c_1, c_2)))}{log(p(c_1)) + log(p(c_2))}$$

- Compares these methods by doing a 'malapropism detection experiment'. Meaning can the methods be used to detect when random words have been altered in a document. Jiang Conrath's method did the best. The Lin and Resnik methods had good recall and so were good at giving possible malapropisms but had worse precision. Leacock-Chodorow had better precision than those too but worse than Jiang Conrath.

- *In summary, this paper discusses several techniques that use WordNet to provide measures of semantic similarity of lexical items. The Jiang-Conrath measure was the most effective measure in the authors experiments.*

4 Comparing Taxonomies for Organizing Collections of Documents

- **Taxonomy** is the practice and science of classification. Many taxonomies have a hierarchical structure, but this is not a requirement. Taxonomy uses taxonomic units, known as taxa (singular taxon).

- Discusses both manually and automatically created taxonomies.

- Manual taxonomies include: Library of Congress Subject Headings (LCSH), WordNet domains, Wikipedia Taxonomy, DBpedia ontology

- Automatically created taxonomies include: **Latent Dirichlet Allocation**, *which we will want to study more in detail*, and a model built on **Wikipedia Link Frequencies**

- Compared the taxonomies expressive power with experiments measuring the *cohesion* and *relatedness*.

  - A cohesive cluster is defined as one in which the items are similar while at the same time clearly distinguishable from items in other clusters. This is evaluated by presenting a human subject a cluster with one element that does not belong. We see how often the subject can determine which element does not belong.
  - Relatedness is defined by two questions: Are the two concepts A and B related? f Yes, then how would you best define the relationship? Is A more specific than B, less specific than B, neither, or don?t know?

- The Wikipedia Link Frequency taxonomy performed very well in the experiments, especially in terms of cohesion.

5 Entity Disambiguation with Freebase

- **Entity disambiguation** is the determination of the *sense* of a word or phrase. E.g. determining if 'The pitcher fell over' refers to a baseball player or a jug of water.
- Introduces Freebase as a tool for disambiguation. It is much bigger than Wikipedia and enjoys a better type taxonomy and more complex schemas. The well structured database is not only convenient for a human to browse, but also very suitable for the machine to use.
- Uses an iterative semi-supervised frame- work to perform entity disambiguation with Freebase.
- Use both generative and discriminative models in our framwork, and find that the discriminative model outper- forms the generative one constantly.

6 A Paper Recommender for Scientific Literatures Based on Semantic Concept Similarity by Ming Zhang, Weichun Wang, and Xiaoming Li

- The paper recommender systems are emerging with the explosive growth of the WWW. McNee et al. mapped the Web of citations between papers into the user-item rating matrix where the paper "votes" for the citations it references.
- Defines a concept graph system which measures both similar clicks AND semantic similarity. But they do not discuss the measure of semantic similarity used.
- They do give a definition of concept similarity (see paper) but it is defined in terms of tags and so perhaps not useful to us.

7 Finding Similar Research Papers Using Language Models & Using Semi-Structured Data for Assessing Research Paper Similarity

- The authors represent a document with a *language model*, a probability distribution over the words in the document. Note how this contrasts with the Vector Space model used in the paper: *Similarity Measures for Text Document Clustering*
- **Language Model**:
  - A document d is then assumed to be generated by a given model D. We want to define the probability that model D generates term w: $\mathcal{D}(w) = P(w|D)$.
  - To compare the model $\mathcal{D}_1$ of document d_1 and $\mathcal{D}_2$ of document d_2 to see how similar the two documents are, we use KL-Divergence:
    - $KLD(\mathcal{D}_1||\mathcal{D}_2) = \sum_w \mathcal{D}_1(w) \log \frac{\mathcal{D}_1(w)}{\mathcal{D}_2(w)}$

  - We define the the language model as:
    * $\mathcal{D}(w) = P(w|D) = \lambda_1 P(w|d) + \lambda_2 P(w|k) + \lambda_3 P(w|a) + \lambda_4 P(w|j) + \lambda_5 P(w|\mathcal{C})$ where k is a keyword, a is the author, j is the journal, $\mathcal{C}$ is whole collection of abstracts.
  - Latent Dirichlet Allocation is used to determine a set of topics discussed in the collection of papers. We add an additional term to the above equation: $\lambda_6 P(w|t)$, where t is a topic.

- Claims to have a publicly available dataset with manually tagged papers, but it is no where to be found.

- The evaluation is done with two measures: *Mean Average Precision* and *Mean Reciprocal Rank*. MAP takes into account the position of every hit within the ranking and is defined by:
  - $MAP = \frac{\sum_{r=1}^{|R|} AvPrec(r)}{|R|}$
  - $AvPrec(r) = \frac{\sum_{i=1}^{n} Prec(i) \times rel(i)}{\text{number of relavent documents}}$
    with $Prec(i)$ the precision at cut off i in the ranking (i.e. the the percentage of the i first ranked items that are relevant) and $rel(i) = 1$ if the item at rank i is a relevant document ($rel(i) = 0$ otherwise).

8 General Notes & Observations

- Most of the approaches I have read for content based document similarity require essentially two definitions: a mathematical representation of the documents and a measure over those representations.

- The two most common representations for documents are a *vector space model* in which a feature vector is used to represent the document and a probabilistic or *language model* in which a probability distribution over the terms in the documents is used as the representation.

- A number of measures can be used to determine the distance between two documents in these models. For the *vector space model* it seems that the *cosine* distance is often used. For the *language model* the probability distributions are usually compared using KL divergence.

- One problem, which is often encountered in information extraction, is the difficulty of two different terms having the same meaning. The most common way to handle this problem is the use of topic clustering by Latent Dirichilet Allocation.

- We could incorporate predicate argument structure in a couple unique ways. For one, we could include it in the language model of *Finding Similar Research Papers Using Language Models & Using Semi-Structured Data for Assessing Research Paper Similarity.*

- We could also include techniques in entity disambiguation instead of or in addition to LDA.

- Also, it doesn't seem like people are really making use of the fact that Wikipedia provides us with a pretty strong taxonomy of scientific topics. This would make an excellent starting point for doing entity disambiguation or LDA.

- I'm sure people have done this, but I haven't found it done yet. We could use the references section of the papers and apply an algorithm like PageRank to do the recommendations/similarity.
- On that note, I think we need to make a decision if we are doing document similarity or if we are doing a recommendation system. They definitely overlap, but I think we should focus on document similarity.

9 Latent Dirichlet Allocation (from Wikipedia)

- Latent Dirichlet Allocation (LDA) is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.
- For example, an LDA model might have topics that can be classified as *CAT_related* and *DOG_related*. A topic has probabilities of generating various words, such as milk, meow, and kitten, which can be classified and interpreted by the viewer as *"CAT_related"*. Naturally, the word cat itself will have high probability given this topic. The *DOG_related* topic likewise has probabilities of generating each word: puppy, bark, and bone might have high probability. Words without special relevance, such as the (see function word), will have roughly even probability between classes (or can be placed into a separate category). A topic is not strongly defined, neither semantically nor epistemologically. It is identified on the basis of supervised labeling and (manual) pruning on the basis of their likelihood of co-occurrence. A lexical word may occur in several topics with a different probability, however, with a different typical set of neighboring words in each topic.
- Each document is assumed to be characterized by a particular set of topics. This is akin to the standard bag of words model assumption, and makes the individual words exchangeable.
- `[TO BE CONTINUED]`. See http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

10 `tf-idf` from Wikipedia

- `tf-idf`, term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.
- The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others.
- Term frequency, `tf` is defined as:
  - $tf(t, d) = 1$ iff t occurs in d
  - Logarithmic or augmented alternatives also exist

- Inverse document frequency, `idf`, is a measure of whether the term is common or rare across all documents. It is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.
  - $idf(t, \mathbf{D}) = \log \frac{|\mathbf{D}|}{\sum_{d \in \mathbf{D}} tf(t,d)}$
- Therefore, `tf-idf` is defined:
  - $tfidf(t, d, \mathbf{D}) = tf(t, d) \times idf(t, \mathbf{D})$
- A high weight in `tf-idf` is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms. Since the ratio inside the `idf`'s log function is always greater than or equal to 1, the value of `idf` (and `tf-idf`) is greater than or equal to 0. As a term appears in more documents, the ratio inside the logarithm approaches 1, bringing the `idf` and `tf-idf` closer to 0.