YTEX Semantic Similarity Measures

Author: Vijay Garla

Date: July 20th, 2012

This document describes the semantic similarity measures implemented by YTEX. We modified some measures so that they conform to the universal definition of similarity presented by Lin [1]: measures are limited to the interval [0,1], and the similarity between a concept and itself is 1.

1. Path finding measures

We focus on the *Path, Leacock and Chodorow* (*LCH*), and *Wu-Palmer* path finding measures. These measures are based on the number of nodes (*path*(c_1, c_2), or *p)* in the shortest path separating two concepts, c_1 and c_2. The shortest path between two concepts traverses their Least Common Subsumer (*lcs*(c_1, c_2)), i.e. their closest common parent. The depth (*depth*(*c*)) of a concept is defined as the number of nodes in the path to the root of the taxonomy; and *d* represents the maximum depth of a taxonomy.

Path defines the similarity between two concepts simply as the inverse of the length of the path separating them [2]:

$$sim_{path}(c_1, c_2) = 1/\ p \tag{1}$$

LCH is based on the ratio of path length to depth, but performs a logarithmic scaling [3]. Originally, LCH was defined as

$$sim_{lch}^{unscaled}(c_1, c_2) = -\log(p/2d) = \log(2d) - \log(p) \tag{2}$$

where *d* represents the maximum depth of the taxonomy. As proposed in [4], we scale LCH to the unit interval by dividing by log(2*d*). Dividing by a constant value has no effect on the spearman correlation with benchmarks: the relative ranks of concept pair similarities remain the same.

$$sim_{lch}(c_1, c_2) = 1 - \frac{\log(p)}{\log(2 \times d)} \tag{3}$$

Wu & Palmer scales the depth of the LCS by the length of the path between two concepts [4]:

$$sim_{wp}^{unscaled}(c_1, c_2) = \frac{2 \times depth(lcs(c_1, c_2))}{path(c_1, lcs(c_1, c_2)) + path(c_2, lcs(c_1, c_2)) + 2 \times depth(lcs(c_1, c_2))} \tag{4}$$

One problem with this definition is that the similarity of a concept with itself is less than 1 (if $c_1 = c_2$, then $path(c_1, lcs(c_1, c_2)) + path(c_2, lcs(c_1, c_2)) = 2$). Instead, we adopt the definition of Wu & Palmer used in the Natural Language Toolkit [5]:

$$sim_{wp}(c_1, c_2) = \frac{2 \times depth(lcs(c_1, c_2))}{p - 1 + 2 \times depth(lcs(c_1, c_2))} \tag{5}$$

Under this definition, if $c_1 = c_2$, then p-1 = 0, and the similarity measure evaluates to 1.

2. IC based measures

Information content can be estimated solely from the structure of a taxonomy (intrinsic IC), or from the distribution of concepts in a text corpus in conjunction with a taxonomy (corpus IC) [6–8].

The corpus IC ($IC_{corpus}(c)$) of a concept is defined as the inverse of the log of the concept's frequency [6]. The frequency of a concept is recursively defined using a taxonomy: it is based

on the number of times the concept *c* occurs within a corpus ($freq(c,C)$), together with the number of times its children occur:

$$IC_{corpus}(c) = -\log(freq(c)) \tag{6}$$

$$freq(c) = freq(c,C) + \sum_{c_s \in children(c)} freq(c_s) \tag{7}$$

We follow the intrinsic IC definition proposed by Sanchez et al [8]:

$$IC_{intrinsic}(c) = -\log\left(\frac{\frac{|leaves(c)|}{|subsumers(c)|}+1}{max\_leaves+1}\right) \tag{8}$$

where *leaves(c)* is the number of leaves (concepts without children) that are descendants of the concept *c*; *subsumers(c)* contains *c* and all its ancestors. The ratio of leaves to subsumers quantifies the information a concept carries– the more leaves a concept has relative to the number of ancestors, the less information it carries; this is normalized to the unit interval by *max_leaves*, the total number of leaves in the taxonomy.

The IC based Lin measure compares the IC of a concept pair to their LCS's IC: the greater the LCS's IC (i.e. the more specific the LCS), the more 'similar' the pair of concepts.

$$sim_{lin}(c_1,c_2) = \frac{2 \times IC(lcs(c_1,c_2))}{IC(c_1)+IC(c_2)} \tag{9}$$

Sanchez & Batet redefined path finding measures in terms of information content [8]. Path finding measures are defined in terms of the path length *p* and the maximum depth *d*. Sanchez & Batet proposed redefining the maximum depth d as ic_{max}, the maximum information content of any concept; and proposed redefining the minimum path length *p* between two concepts in terms of Jiang & Conrath's semantic distance [8], [9] :

$$dist_{jc}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times IC(lcs(c_1, c_2)) \quad (10)$$

The IC-based *LCH* measure is obtained simply by substituting $dist_{jc}$ and ic_{max} for p and d in equation 3 (1 is added to $dist_{jc}$ to avoid taking the logarithm of 0):

$$sim^{*}_{lch\_ic}(c_1, c_2) = 1 - \frac{\log\left(dist_{jc}(c_1, c_2) + 1\right)}{\log(2 \times ic_{\max})} \quad (11)$$

One problem with this definition is that the IC-based LCH can assume negative values. We modify this as follows:

$$sim^{*}_{lch\_ic}(c_1, c_2) = 1 - \frac{\log\left(dist_{jc}(c_1, c_2) + 1\right)}{\log(2 \times ic_{\max} + 1)} \quad (12)$$

Both Sanchez & Batet's and our definitions of the IC-based LCH are monotonically decreasing functions of $dist_{jc}$, and thus produce identical spearman correlations with benchmarks.

The IC-based *Path* measure (als known as the Jiang & Conrath *similarity* measure) is obtained simply by substituting $dist_{jc}$ for p (1 is added to $dist_{jc}$ to avoid dividing by 0):

$$sim_{path\_ic}(c_1, c_2) = \frac{1}{dist_{jc}(c_1, c_2) + 1} \quad (13)$$

3. References

[1] D. Lin, "An Information-Theoretic Definition of Similarity," in *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998, pp. 296–304.

[2] T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, and C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," *J Biomed Inform*, vol. 40, no. 3, pp. 288–299, Jun. 2007.

[3] C. Leacock and M. Chodorow, "Combining local context with WordNet similarity for word sense identification," in *WordNet: A Lexical Reference System and its Application*, 1998.

[4] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, Las Cruces, New Mexico, 1994, pp. 133–138.

[5] S. Bird, E. Loper, and E. Klein, “NLTK Toolkit.” [Online]. Available: http://nltk.googlecode.com/svn/trunk/doc/api/nltk.corpus.reader.wordnet-pysrc.html#Synset.wup_similarity. [Accessed: 08-Jun-2012].
[6] P. Resnik, “Using Information Content to Evaluate Semantic Similarity in a Taxonomy,” in *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 448–453.
[7] N. Seco, T. Veale, and J. Hayes, “An Intrinsic Information Content Metric for Semantic Similarity in WordNet,” in *ECAI’2004, the 16th European Conference on Artificial Intelligence*, 2004.
[8] D. Sánchez and M. Batet, “Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective,” *Journal of Biomedical Informatics*, vol. 44, pp. 749–759, Oct. 2011.
[9] J. J. Jiang and D. W. Conrath, “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,” *CoRR*, p. -1–1, 1997.