Partial Measure of Semantic Relatedness based on the Local Feature Selection

Maciej Piasecki and Michał Wendelberger

Institute of Informatics, Wrocław University of Technology, Poland
maciej.piasecki@pwr.edu.pl

Abstract. A corpus-based Measure of Semantic Relatedness can be calculated for every pair of words occurring in the corpus, but it can produce erroneous results for many word pairs due to accidental associations derived on the basis of several context features. We propose a novel idea of a *partial measure* that assigns relatedness values only to word pairs well enough supported by corpus data. Three simple implementations of this idea are presented and evaluated on large corpora and wordnets for two languages. Partial Measures of Semantic Relatedness are shown to perform better in tasks focused on wordnet development than a state-of-the-art 'full' Measure of Semantic Relatedness. A comparison of the partial measure with a globally filtered measure is also presented.

1 Introduction

Measures of Semantic Relatedness (henceforth, MSRs) built within Distributional Semantics are one of ways to mine for lexical semantic knowledge out in large text corpora. The extracted knowledge can be utilised in wordnet development, cf [10]. MSR assigns a numerical value to a word pair specifying their semantic relatedness on the basis of their corpus distribution and contextual *features* characterising word occurrences, e.g. a feature is the number of co-occurrences with a word or a lexico-syntactic structure in a specified text context. MSRs construction can be tuned to different needs [1]. For instance, in wordnet development, we expect that words linked by lexico-semantic relations are assigned higher values by MSR.

MSR can be used as knowledge source for wordnet expansion algorithms, e.g. [14,13], but it can also be directly consulted by lexicographers. In both cases, the basic information we want to obtain from MSR is a list of k words that are most semantically related to the given word and such a list is called here a *k-list*. It can be thousands words long for a word, but only the top part is useful in both applications. K-lists can include many errors of two main kinds: words associated by plenty of other relations than those used in wordnets and words linked completely accidentally. The latter errors are especially characteristic for less frequent words. For instance an MSR built on British National Corpus (BNC)[1] (Sec. 3) produces for *frost* (relatedness values in brackets): ***rain*** (0.151), *forsbrand* (0.15), *gale* (0.146), *hawksworth* (0.124), *fitzwilliam* (0.122), ***sleet*** (0.114) ***snow*** (0.11) *rime* (0.1), ***dew*** (0.1), *buda* (0.1), *lawley* (0.1), ***fog*** (0.096), …While links of non-wordnet relations can be helpful in wordnet expansion

[1] http://www.natcorp.ox.ac.uk

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 336–343, 2014.

– e.g. to characterise semantic fields – the accidental associations are problematic for both lexicographers and algorithms (except those that express idiosyncratic properties).

A simple but commonly used method is a threshold-based filtering of words and features during MSR construction, e.g. MSR computation only for those words that are frequent enough in the corpus or have weight value high enough, e.g. [9,15,11,1]. Such global filtering is a radical solution and results in information loss. In relation to particular words, the global frequency threshold can be too high or too low, e.g. some words occurring around 100 times in the 1 billion corpus can still have good description. A large wordnet mostly consists of less frequent words and wordnet expansion is done mostly for words that are less frequent even in a huge corpus.

The strength of association between two words depends on features shared by them. Some words can be mistakenly associated due to accidental features, e.g. an MSR associated *blue spruce* with *antler* because both were described by adjectives *branchy* and *towerring*, and only a few features more. For such words we should decide whether they are semantically related on the basis of the collected information. Our idea is to recognise word pairs for which we have enough amount of the corpus-based information characterising their relationship and compute MSR only for them. The goal is to create a method which will identify words that can be compared and for which the MSR value can be calculated. In sum, we aim at *a partial MSR* that assigns values only to those word pairs for which enough information was collected from the corpus.

2 Method

Two words should be assigned a higher value by an MSR only if they share *a large enough number* of good *quality* features. However, if they do not, we cannot assess their semantic association, as the corpus data are always partial. Thus, when the data supporting a word pair association are too limited, MSR should abstain and does not assign a value to it.

Good quality features must provide enough information to support semantic association of the compared words. The key issues are: how to measure feature quality and how many features must be shared? Statistical association measures or information theoretic measures are commonly used to weigh descriptions provided by features for individual words. For instance, Pointwise Mutual Information (PMI) was often applied and reported to express good performance, e.g. [9,13,2]. However, PMI overestimates some features, especially for less frequent words and there is no universal threshold for PMI values that guarantees appropriate feature selection, i.e. it is clear that only positive PMI values should be used but there are no further thresholds. Most statistical association measures can be misinterpreted in the case of infrequent words. Thus, instead of filtering procedures working on the global scale, we propose a general scheme of *a partial MSR computation* that can be instantiated with different specific solutions. We identify *globally unimportant* features as those that do not discriminate among different words, e.g. *e*% of features with the highest entropy in the matrix, and *locally unimportant* features that according to the assumed *measures of feature importance* cannot be treated as an important part of the word description.

Let: **M** be a coincidence matrix of words and features, C_E – a set of globally unimportant features, σ – a matrix raw similarity function, x, y – words, R_x, R_y – unweighted row vectors (of frequencies), and W_x, W_y – weighted row vectors.
A partial MSR for x and y is calculated in two steps. First, locally important features for x and y are identified. Next, the partial MSR value is calculated only if the features shared between x and y fulfill the specified conditions.

1. For each $i \in x, y$ the set of locally important features LF_i is defined as:
   (a) LF_i = all non-zero features from R_i minus features from C_E
   (b) $LF_i = f_{imp}(LF_i, R_i, W_i)$
2. Similarity computation:
   (a) If $f_{part}(LF_x, R_x, W_x, LF_y, R_y, W_y)$ equals **true**
   then $\text{MSR}(x, y) = \sigma(W_x, W_y)$ else **unknown**.

The functions: f_{imp} for filtering out locally unimportant features and f_{part} for deciding about sufficient amount of information for comparison, are parameters.
Simple frequency-based partial MSR is based on a simple heuristics originating from the manual inspection of a sample of k-lists (later excluded from tests). Single co-occurrences of a feature and a word are often accidental, so f_{imp} returns only features j such that $R_i[j] > 1$.

Next, we observed that two 'good' features are mostly enough to make the MSR value meaningful, so $f_{part} =$

1. If $|LF_x \cap LF_y| > 2$ then return **true**.
2. If $\exists c.(LF_x \cap LF_y| = \{c\}$ and
   – $\exists c' \in LF_x.(c' \neq c \wedge R_y[c'] > 0)$
   – or $\exists c'' \in LF_y.(c'' \neq s \wedge R_x[c''] > 0)$ then return **true**, else **false**.

So, according to the simple heuristics MSR value is calculated only for words sharing at least two features co-occurring with each of them at least twice. These criteria are applied in parallel to the weighting of features and similarity computation. In all experiments we used PMI weighting and cosine similarity.
PMI-based partial MSR – is based on the assumption that features (W_i) are weighted by popular and well-performing PMI. On the basis of the previous experience, we selected Lin's version of PMI [9] and the cosine similarity measure, cf [13,2]. Lin used co-occurrences of words and features as events:

$$\text{PMI}(w, f) = \log\left(\frac{c(w, f) \sum_{w', f'} c(w', f')}{\sum_{f'} c(w, f') \sum_{w'} c(w', f)}\right)$$

The aim was to exchange frequency criteria to PMI-based criteria, but zero seemed to be too weak and higher values are not theoretically motivated. On the basis of the manual inspection of the same date sample, we set the PMI threshold for locally important features to ≥ 1.0 (the global threshold was kept on 0). In addition, we wanted to favour features grouping words into semantic classes or shared among words of the same class. Broad semantic classes were identified on the basis of the wordnet hypernymy structure. Starting from the top synsets hyponymic subtrees were iteratively divided into smaller

ones. In each step the largest hypernymy subtree was selected and divided into two subclasses. The direct and indirect hypernyms of the selected subtree were included into both created subclasses. The process ended when the predefined number of subclasses was reached.

Globally unimportant features were defined as 1% with the highest entropy in relation to words and 5% of the highest entropy to semantic classes. Class-base coincidence matrix was constructed on the basis of the word-based matrix. Occurrences of the polysemous words were added to all their classes.

The f_{imp} function was re-defined on the basis of the PMI threshold:
f_{imp} filters out all features j such that $W_i[j] \leq 1$
and used in the function f_{part}, which stayed unchanged.

Hypernymy-based partial MSR – explores more the developed semantic noun classification. The criterion for locally important features was not changed, because semantic classes of words outside wordnet are not known. However, the associations between features and semantic classes can be discovered on the basis of a large wordnet and a large corpus. The function f_{part} was updated to promote words linked by features of the same class[2]:

1. If $\exists A \in Classes.$
   $|\{f : f \in (LF_x \cap LF_y) \wedge desc(f, A)\}| \geq k$ then return **true**.
2. If $|LF_x \cap LF_y| > n$ then **true** else **false**.

As the first condition is to strict alone, MSR value is also calculated for words sharing at least n locally important features, but we intend to have $n \geq k$.

3 Evaluation

Several approaches to the evaluation of MSRs were proposed, e.g. a comparison to human decisions [16], solving tests similar to TOEFL [6,4,12], and comparison with the wordnet-based similarity [9,15]. However, for the purposes of wordnet development, the most important is to have possibly many instances of wordnet relations in the top of the k-list, in our test:

- for each test word x, a set of all words connected to it in the wordnet is generated,
- and the set is compared with the k-list of x generated by the MSR.

Cut-off *precision* for k-lists is defined as $|L_M(k)|/|L_W|$ and *recall* as $|L_M(k) \cap L_W|/|L_W|$, where $L_M(k)$ is a set of word pairs $\langle x, y \rangle$ such that y belongs to k-list of x, and L_W is a set of all pairs extracted from the wordnet for the test words. The wordnet set includes: synonyms, direct and indirect hypo/hypernyms (up to 3 links), cousins (up to 2 hypernymic and hyponymic links), mero/holonyms and the words linked directly by lexical relation. In the case of polysemous words all synsets were considered.

In addition, we also applied the second evaluation method based on testing MSR ability to predict wordnet-based semantic similarity [15]. Correlation between k-lists

[2] In practice, due to the polysemy of the words in the original matrix most features are associated with many semantic classes.

Table 1. Tests on British National Corpus: k – No of positions for the cut-off precision P and recall R, *Cor.* – the correlation with the similarity measure on WordNet 3.1; *Cov* – ratio between the No pairs extracted by a MSR and the full MSR, *Hits* – No of word pairs in the k-best lists of the MSR.

Func.	k	P [%]	R [%]	F1 [%]	Cov. [%]	Hits	Cor.
Full	**5**	11.42	0.03	0.06	100.00	9 926	0.003459
	10	10.87	0.05	0.11	100.00	18 884	0.000797
	20	10.40	0.10	0.20	100.00	36 102	0.000164
	50	9.77	0.24	0.47	100.00	84 499	0.000018
Full globally filtered	**5**	14.37	0.04	0.07	96.52	12 054	—
	10	13.58	0.07	0.13	96.60	22 786	—
	20	12.69	0.12	0.25	96.71	42 596	—
	50	11.54	0.28	0.55	96.96	96 791	—
Simple	**5**	**31.09**	**0.07**	**0.15**	70.04	18 931	**0.003901**
	10	**29.07**	**0.14**	**0.27**	68.42	34 553	**0.000899**
	20	**27.03**	**0.24**	**0.48**	66.07	61 978	**0.000186**
	50	**24.09**	**0.50**	**0.98**	61.51	128 175	**0.000021**
PMI-based	**5**	24.32	0.06	0.12	**93.26**	19 719	0.003741
	10	23.23	0.11	0.22	**92.05**	37 155	0.000867
	20	22.20	0.21	0.41	**90.03**	**69 355**	0.000179
	50	20.36	0.45	0.88	**85.80**	**151 065**	0.000020
Hypernymy-based	**5**	24.33	0.06	0.12	93.25	**19 725**	0.003766
	10	23.27	0.11	0.22	92.00	**37 186**	0.000874
	20	22.23	0.21	0.41	89.87	**69 335**	0.000180
	50	20.41	0.45	0.88	85.38	150 722	0.000020

(treated as ranking lists) produced by an MSR and a wordnet-based similarity measure was analysed. Following [15], we used the Jiang and Conrath similarity measure [5] (JC measure) to generate wordnet-based k-lists for test words. For ranking list comparison, neighbour set comparison technique proposed by [8] and adapted in [15] for this task was applied:

$$Cor(S, S') = \frac{\sum_{w \in S \cap S'} (w, S)(w, S')}{\sum_{i=1}^{k} i^2}$$

where S, S' are two k-lists (ranking lists) and (w, S) returns $k - ranking(w, S)$.

Experiments were performed on the two world largest wordnets: Princeton WordNet 3.1 (PWN) [3] and plWordNet 2.1 – the Polish wordnet (plWN) [10], which is not translated from PWN and has also a slightly different character. For the experiments on English, we used BNC, which has been often used for building MSRs. For Polish, we have built a joint corpus of 1.8 billion tokens by merging together freely available corpora and larger texts acquired from Internet. BNC was processed by MiniPar dependency parser [7], and the Polish joint corpus with morpho-syntactic constraints from [13]. As a result English, and Polish words were described by lexico-syntactic relations used as features in the matrices.

The baseline *full MSRs* were built for all one-word nouns from both corpora. All possible features were collected initially, but first 1% features with the highest entropy.

Table 2. Tests on the Polish joint corpus (the same labels as in Tab. 2)

Func.	k	P	R	F1	Cov. [%]	Hits	Cor.
Full	**5**	24.80	0.11	0.22	100.00	62 239	–
	10	22.85	0.21	0.41	100.00	114 473	–
	20	20.96	0.38	0.75	100.00	209 401	–
	50	18.49	0.82	1.56	100.00	450 468	–
Full globally filtered	**5**	29.26	0.13	0.26	95.13	69 843	0.01124
	10	26.96	0.24	0.49	95.30	128 696	0.002585
	20	24.62	0.45	0.88	95.56	235 049	0.000540
	50	21.34	0.97	1.85	97.88	508 680	0.00006
Simple	**5**	**40.08**	**0.18**	**0.35**	79.08	79 535	0.011253
	10	**36.45**	**0.32**	**0.63**	78.51	143 353	0.002588
	20	**32.72**	**0.57**	**1.11**	77.60	253 646	0.000541
	50	**27.55**	**1.15**	**2.21**	77.15	517 699	0.00006
PMI-based	**5**	36.38	0.16	0.32	**90.92**	**82 990**	**0.011254**
	10	33.59	0.30	0.59	**90.29**	**151 934**	**0.002589**
	20	30.73	0.54	1.06	**89.31**	**274 222**	**0.000541**
	50	26.56	1.13	2.17	**89.04**	**576 006**	**0.00006**
Hypernymy-based	**5**	36.39	0.16	0.33	90.89	82 987	**0.011254**
	10	33.61	0.30	0.59	90.26	151 936	**0.002589**
	20	30.74	0.54	1.06	89.26	274 185	**0.000541**
	50	26.58	1.13	2.17	88.92	575 810	**0.00006**

So, the English matrix was initially of the size: 39,411 nouns and 124,830 features, and 58,781 nouns and 643,894 features for Polish. Next, the influence of the *global filtering* was tested. In addition to the entropy threshold, minimal frequency for words was set to 5, for features to 20, and all features with less than 20 non-zero cells after PMI weighting were removed from the matrix. Concerning parameter values, *simple partial MSR* (PMSR) was used as described earlier. For *PMI-based PMSR*, we set f_{imp}: 200 semantic classes, minimal PMI value 1, top entropy features vs words 1%, top entropy features vs classes 5%, and f_{part} unchanged. In hypernymy-based PMSR, f_{part} was instantiated to $k = 2$ and $n = 2$. In all cases global feature frequency threshold was 5, entropy threshold was 1%, and cosine similarity was used.

The evaluation results are presented in Tab. 1 for English and in Tab. 2 for Polish. All partial MSRs have higher precision than the 'full' MSRs and the globally filtered MRSs, i.e. mostly non-related words were filtered out. Proper words are sometimes eliminated too, but, surprisingly, partial MSRs also have better recall. This is caused by lifting up proper words from lower ranking positions to the k top words. The difference between all partial MSRs and both types of the 'full' MSRs is statistically significant, with the confidence level higher than $1 - 0.005$ (measured by t-test applied to average results on 50 random samples) in both types of evaluation. Only some differences between partial MSRs are statistically significant with the confidence $1 - 0.05$.

Coverage – a ratio of the number of word pairs generated by partial and full MSR, is higher in the case of PMI-based and hypernymy-based PMSRs. They generate more word pairs. Some of them can be proper, but not covered by wordnets. Both PMSRs

have the highest numbers of hits, i.e. they produce longer *k*-lists on average that cover more wordnet relation instances.

The following examples illustrate changes introduced to *k*-lists by a simple PMRS. For the English word *frontier* the full MSR produces: ***border*** (0.108), *euphrates* (0.104), *mindanao* (0.0838) *findhorn* (0.0791), *memnon* (0.0755), *negus* (0.0709), ***province*** (0.0686), *demerit* (0.0681) *eurythmics* (0.0661), *weinstock* (0.0661) *vase-painting* (0.0655), ***ambassador*** (0.0639) *pattaya* (0.0638) *non-russian* (0.0626), ***siberia*** (0.0617), *reik* (0.0615), ***territory*** (0.0603), ***coast*** (0.0602), *tunney* (0.0596), *quad-fx* (0.0584), *gondwana* (0.0554), *descendent* (0.0553), *beowulf* (0.0551), ***national*** (0.0548), ***sovereignty*** (0.0548). The simple partial MSR filters out all associations from this list except the ones in bold.

The MSR built for Polish provides better description for many words, as the joint corpus is more than 18 times bigger than BNC. However, less frequent words can cause accidental associations, e.g. for the Polish noun *kapsuła* 'capsule' the full MSR generates: ***właz*** 'hatch' (0.1011), ***statek kosmiczny*** 'spacecraft' (0.0816), *konserwatornia* '≈restoration workshop' (0.0803), *izostazja* 'isostasy' (0.078), ***luk*** 'hatchway' (0.0766), *odcumowanie* 'unmooring' (0.075), ***kabina*** 'cabin' (0.071), ***wahadłowiec*** 'shuttle' (0.0696), ***moduł*** 'module' (0.0687), *smażalnik* '≈machine for frying' (0.0685), *odżelaziacz* '≈iron remover' (0.0685), ***zasobnik*** 'tray' (0.0677), *sferolit* 'spherulite' (0.0668), *łuszczynka* 'siliqua' (0.0661), *ser ementalski* 'Emmental cheese' (0.066), ***prom*** 'space shuttle' (0.0653), …

4 Conclusions

A novel idea of the partial MSR was proposed that assigns semantic relatedness values only to word pairs that are enough well described by the corpus-based information. A partial MSR seems to be a better knowledge source for wordnet development and lexicography than a 'full' MSR. Simple but effective implementations of partial MSRs were tested on the two world largest wordnets, and two corpora: a medium and huge one and two very different languages. All evaluated variants are language and resource independent. An obvious drawback of a partial MSR is that it does not assign values to some word pairs. However, a full MSR often assigns unreliable values that are not well supported by corpus data.

Acknowledgments Partially financed by the Polish Ministry of Science and Higher Education as a Investment in CLARIN-PL Research Infrastructure.

References

1. Baroni, M., Lenci, A.: Distributional memory: A general framework for corpus-based semantics. Computational Linguistics 36(4), 637–721 (December 2010)
2. Bullinaria, J.A., Levy, J.P.: Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. Behav Res Methods 44(3), 890–907 (Sep 2012)
3. Fellbaum, C. (ed.): WordNet – An Electronic Lexical Database. The MIT Press (1998)

4. Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R., Wang, Z.: New experiments in distributional representations of synonymy. In: Proc. of the 9th Conf. on Computational Natural Language Learning. pp. 25–32. ACL, Ann Arbor, Michigan (2005)
5. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the International Conference on Research in Computational Linguistics (ROCLING X). Taiwan (1997)
6. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition. Psychological Review 104(2), 211–240 (1997)
7. Lin, D.: Principle-based parsing without overgeneration. In: Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (1993)
8. Lin, D.: Using syntactic dependency as local context to resolve word sense ambiguity. In: Proc. of the 35th ACL and 8th EACL. pp. 64–71. ACL, Madrid (1997)
9. Lin, D.: Automatic retrieval and clustering of similar words. In: Proc. of the 35th ACL and 17th Inter. Conf. on Computational Linguistics. pp. 768–774. ACL (1998)
10. Maziarz, M., Piasecki, M., Rudnicka, E., Szpakowicz, S.: Beyond the transfer-and-merge wordnet construction: plWordNet and a comparison with WordNet. In: Proc. of the Inter. Conf. Recent Advances in Natural Language Processing RANLP 2013, Hissar, Bulgaria. INCOMA Ltd. and ACL (2013)
11. Navigli, R., Velardi, P., Faralli, S.: A graph-based algorithm for inducing lexical taxonomies from scratch. In: Proceedings of IJCAI (2011)
12. Piasecki, M., Szpakowicz, S., Broda, B.: Extended similarity test for the evaluation of semantic similarity functions. In: Vetulani, Z. (ed.) Proce. of the 3rd Language and Technology Conference, Poznań. pp. 104–108 (2007)
13. Piasecki, M., Szpakowicz, S., Broda, B.: A Wordnet from the Ground Up. Oficyna Wydawnicza Politechniki Wrocławskiej (2009), `http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf`
14. Snow, R., Jurafsky, D., Ng, A.Y.: Semantic taxonomy induction from heterogenous evidence. In: Proc. of the Joint Conf. of the International Committee on Computational Linguistics and ACL. pp. 801–808 (2006)
15. Weeds, J., Weir, D.: Co-occurrence retrieval: A flexible framework for lexical distributional similarity. Computational Linguistics 31(4), 439–475 (2005)
16. Zesch, T., Gurevych, I.: Automatically creating datasets for measures of semantic relatedness. In: Proceedings of the Workshop on Linguistic Distances. pp. 16–24. Association for Computational Linguistics, Sydney, Australia (2006)