Fuzzy Synsets, and Lexicon-Based Sentiment Analysis

Sayyed-Ali Hossayni[1 a,b], Mohammad-R Akbarzadeh-T[b], Diego Reforgiato Recupero[c,e], Aldo Gangemi[d,c], Josep Lluís de la Rosa i Esteva[a]

[a] Agents Research Lab, TECNIO Centre EASY, University of Girona, Girona, Catalonia, Spain.
[b] SCIIP center of excellence, Ferdowsi University of Mashhad, Mashhad, Iran.
[c] STLab, Institute of Cognitive Sciences and Technologies (ISTC), CNR, Italy
[d] Laboratoire d'Informatique de Paris Nord, Université Paris 13 - CNRS, France
[e] Department of Mathematics and Computer Science, University of Cagliari, Italy
hossayni@iran.ir; akbazar@um.ac.ir; diego.reforgiato@unica.it;
aldo.gangemi@lipn.univ-paris13.fr; peplluis@silver.udg.edu

Abstract. One of the widely used approaches to Sentiment Analysis (SA) is lexicon-based approach that depends on sentiment-annotated lexical resources (such as SentiWordNet (SWN)). A broad variety of such resources are Synset-based Lexical Databases (SLDs) (e.g. SWN is based on WordNet (WN)) and represent sentiment degrees of synonym groups of LDs, called "synsets." However, synsets themselves were open to criticism because although, in reality, not all the members of a synset represent its meaning with the same degree, in SLDs, they are, identically, considered as members of their synset. Therefore, the fuzzy version of synsets was proposed in a small number of previous studies. Fuzzy synsets can upgrade such lexicon-based SA by which the future SA systems can discriminate between word-senses of a same synset, how much each of them contains the sentiment load of that synset. But, to the best of our knowledge, none of the studies on fuzzy synsets has proposed any algorithm for providing fuzzy versions of "predefined synsets" of an SLD. In this study, we present the idea of an algorithm for constructing fuzzy version of any SLD of any language, given a corpus of that language and a word-sense-disambiguation system of that language/SLD.

Keywords: Sentiment analysis, lexicon-based approach, Synsets, Fuzzy synsets, Probability to possibility transformation

1 Introduction

Sentiment Analysis (SA) has received broad attention in the recent decade. However, extracting sentiment information from unstructured text data is a multi-disciplinary problem, considering that sentiments can be expressed in numerous forms and combinations where it might be difficult to find any sort of regular behavior.
From one point of view, the majority of approaches to SA are divided into two categories: "Machine learning approach," and "Lexicon-based (LB) approach" [21]. The

[1] Corresponding author

former utilizes Machine Learning algorithms mainly to solve SA as a regular text classification problem using syntactic and/or linguistic features, whereas the latter basically utilizes an opinion lexicon (i.e. a list of opinion words and phrases), and a set of rules for determining the opinions orientations in a sentence and also considers opinion shifters and but-clauses [20]. The former provides maximum accuracy whereas the latter provides better generality [26]. However, "Lexicon-based approach is more often used recently"[2] [21]. LB approach (utilizing opinion lexicons [21] as well as generating them [20] for SA purposes) is further divided into dictionary-based and corpus-based categories. In the former, the domain of the opinion-words is as wide as the domain of a complete dictionary, whereas in the latter the domain is limited to those included in the analyzed corpus (corpora). The corpus-based approach, alone, is not as effective (for identifying all opinion words) as the dictionary-based approach because it is hard to prepare a huge corpus to cover all the English words. Conversely, the corpus-based approach has the major advantage of finding domain- & context-specific opinion words and their orientations using a domain corpus [20].

In brief, based on the [20] [21] categorizations, LB-SA approaches are categorized to dictionary-based and Corpus-based the latter of which has the sub-approaches of Statistical, Semantic, and NLP[3]-based. Synset-based Lexical databases (SLDs) such as WordNet (WN) [14] that organize words of a language in synonym groups -called synsets-[4] are being utilized by dictionary-based approach as well as semantic sub-approach of the corpus-based approach in SA, several of which take advantage of the synset-based opinion lexicons such as SentiWordNet (SWN) [13][2]. (SWN is a lexical resource in which each WN synset is associated to Objective, Positive, and Negative values in the continuous interval [0,1] for describing how objective, positive and negative the terms contained in that synset are). However, in the prevalent SLDs such as WN all the members of a synset are supposed to belong to a synset with a same degree and convey the meaning of that synset at a same level. In other words, such SLDs assume synsets to be crisp and non-fuzzy sets. This simple assumption does not always properly model the complex nature of "meaning" in natural languages. For example, consider the following synset of the WN: *Synset('flower.n.02'): {flower, bloom, blossom}*. Flower, bloom, and blossom are each addressing one of the word-senses of the Synset('flower.n.02'). Upon WN information, this synset contains the word-senses with the meaning "reproductive organ of angiosperm plants especially one having showy or colorful parts"; but, obviously, the compatibility of its three word-senses with its definition is not the same that might be considered as a drawback for such SLDs. In the next section, we address a new generation of synsets, fulfilling this drawback.

---

[2] There is a bit of modification in the phrase, regarding that it is a part of a larger sentence.

[3] Natural Language Processing

[4] WN [14] is an LD for the English language that in addition to grouping English words into synsets, provides short definitions, usage of examples of the synsets, and a number of relations among those synsets and their members.

2 Fuzzy synsets, a more informative version of synsets

As mentioned above, usually, it is not the case that compatibility of the word-senses of a synset with the meaning of that synset is in the same degree. It is the reason for which the concept of fuzzy synsets was born. Since 2005, some studies have been conducted where a synset is considered a fuzzy set. In 2005, Veldall [28] , without using the term "fuzzy synset" (even without using the term "synset"), proposed an algorithm for creating fuzzy semantic classes[5] (i.e. synsets) and stated that "different words can represent more or less typical instances of a given concept. Some words may represent clear-cut instances of a given category, while others represent peripheral or border-line cases we let a membership value represent the degree of typicality or compatibility that a word holds toward the concept a class[6] expresses." In 2010, Borin et al. [6] who, to the best of our knowledge, coined the term "fuzzy synsets," viewed them from a pure linguistics point of view, and based them on "synonymy avoidance" [17] concept: *"There is a postulated universal linguistic principle of (full) synonymy avoidance* [7]*. his being an intrinsic characteristic of human language… a dictionary whose fundamental organization is based on the notion of synonymy almost by definition cannot present a faithful reflection of our lexical knowledge, at least not from a linguistic point of view. WN synonyms, as originally defined, should be interchangeable in some contexts, but not necessarily in all contexts* [24]*; in fact, even one context is enough* [1]*. This indicates that synonymy in the WN sense may not correspond exactly to how linguists and lexicographers understand this term, and further that it may be a matter of degree."* In the mentioned study, Borin et al. [6][5] utilized Synlex [18] and SALDO [4] Swedish lexical resources by which they presented an algorithm to create the Swedish fuzzy synsets. In 2011, Gonçalo and Gomes [15] looked at fuzzy synsets from a linguistics point of view expressing that *"from a linguistic point of view, word senses are not discrete and cannot be separated with clear boundaries* [19] [16][7]*. Sense division in dictionaries and lexical resources are most of the times artificial... A more realistic approach for coping with this fact is to represent synsets as models of uncertainty, such as fuzzy sets."* They [15] applied their algorithm on the Portuguese language and proposed Portuguese fuzzy synsets. However, all the mentioned studies have a missing link for being able to upgrade LB-SA systems. That missing link will be addressed in the following section.

3 Fuzzy synset-based lexical databases and upgrading lexicon-based sentiment analysis

In the previous section, we mentioned the drawback of crisp synsets. This drawback also permeates synset-based SA methods including SLD-utilizing LB-SA methods[8],

[5] He applied his algorithm on Norwegian language.
[6] i.e. synset
[7] the original reference was older version of [16]
[8] There are also other synset-based SA methods to which we do not address in this short paper.

because they use the same crisp synsets. For instance, SWN 3.0 assigns a sentiment pair (positive, negative) to each of the WN synsets and assumes all of its word-senses to have the same sentiment load. Such LB-SA methods can be upgraded by fuzzy versions of their utilized crisp synsets, discriminating between word-senses of one fuzzy synset, how much each of its word-senses contains the sentiment load of that fuzzy synset, and thus, assigning a low (high) semantic load to low (high) membership-graded word-senses of that synset. For example, the Synset('run_into.v.01') is annotated as (+0, -0.25) in SWN 3.0. Suppose the fuzzy version of this synset to be {(run_into, 1.0), (encounter, 0.4)}. Then, considering that the word-sense 'encounter' is not fully compatible with this synset (40% compatible), it is not precise to assign (+0, -0.25) (the sentiment load of that synset) to this word-sense in SA process. Its sentiment load does not inherit all the negativity of its synset; yet, it might inherit sentiment of other synsets to which it is compatible (e.g. 'run_into' is also word-sense of Synset('run_into.v.02'), Synset('hit.v.02'), and Synset('meet.v.01')) regarding which upgraded SA methods shall use "graded word-sense assignment" [12][11] and/or fuzzy WSD [27][10] and specify the grade by which 'run_into' belongs to the other 3 synsets and then aggregate the semantic load of all those synsets based on the membership (intra-synset) and grade (inter-synset) of 'run_into' to each of those synsets. Then, the aggregated value would be more informative than simply using (+0.0, -0.25) for it, inheriting from its synset. But, to the best of our knowledge, none of the few studies towards fuzzy synsets have proposed any algorithm for constructing fuzzy version of an SLD, converting its synsets from crisp sets to fuzzy sets, specifying membership-degree of their members (word-senses). Thus, for the mentioned upgrade in SLD-utilizing LB-SAs, an algorithm is yet required for converting the synsets of the existing SLDs to a fuzzy version.

4 Idea of a language-free algorithm for providing fuzzy synsets

In this study, we propose an idea for providing fuzzy version of synsets for predefined synsets: Consider a large-enough corpora of documents of one language; based on the relative frequency of a word-sense of an arbitrary synset 's' (of that language) to the frequency of other word-senses of 's', in the corpus, we can extract the probability of utility of that word-sense among other word-senses of 's'. Then, we can convert those probabilities to possibility values[9] by means of the probability to possibility transformations, proposed by Prade and Dubois in 1983 [9] and 1993 [8]. Then, based on the definition that Zadeh has provided from possibility, in his paper while proposing the possibility theory [30], we can conclude that the extracted possibility values are the same as the membership degrees of the word-senses of the corresponding synset.

By means of this method, we can provide a language-free algorithm for assigning membership functions to synsets of any LD (WN or any other). The only required input of the algorithm, resulting from the suggested idea, will be a large-enough corpus of documents of the opted arbitrary-language (big enough so that relative fre-

[9] By possibility, we are addressing the concept that is subject of Possibility Theory, proposed by Zadeh in 1978 [30].

quency of word-senses can provide trustable probability values) and a precise Word Sense Disambiguation (WSD)[10] system (trustable so that the frequency of word-senses will be real frequencies and not false-detections of word-senses).

5 Conclusion and future works

In this study, we discussed on the potential of the fuzzy synsets in upgrading the synset-based lexicon-based Sentiment Analysis. We highlighted the lacking of an algorithm for generating fuzzy version of predefined synsets in any synset-based lexical database (SLD) (e.g. WordNet (WN)), and suggested an idea of a language-free algorithm that provides fuzzy versions of synsets of any SLD, given a large corpus of documents of the corresponding language and a Word Sense Disambiguation (WSD) system associated with that SLD. In the conference version of this study, we extend the idea, come up with the corresponding algorithm, and also apply it on the English language using the open American national corpus (OANC) and UKB (a well-known graph-based WSD), for constructing fuzzy synsets of English language based on WN.

6 Acknowledgments

AGAUR res. grant 2013 DI 012; MARIO EU Proj.; IDENTITY– n.690907 H2020-MSCA-RISE-2015; QWAVES- RTC-2014-2576-7; ANSwER- RTC-2015-4303-7; CSI-2014 SGR 1469; and NIR-VANA– n. 681787-2 H2020-INNOSUP-2015-2.

7 References

1. Alonge, A. et al.: The Linguistic Design of the EuroWordNet Database. EuroWordNet: A multilingual database with lexical semantic networks. pp. 19–43 Springer Neth. (1998).
2. Baccianella, S. et al.: SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). pp. 2200–2204 (2010).
3. Borin, L.: Mannen är faderns mormor: Svenskt associationslexikon reinkarnerat. LexicoNordica. 12, 39–54 (2005).
4. Borin, L., Forsberg, M.: All in the family: A comparison of SALDO and WordNet. In: Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies. NEALT Proceedings Series. (2009).
5. Borin, L., Forsberg, M.: Beyond the synset: Swesaurus–a fuzzy Swedish wordnet. In: Proceedings of the symposium: Re-thinking synonymy: semantic sameness and similarity in languages and their description. (2010).

[10] In cognitive and computational linguistics, Word Sense Disambiguation (WSD) is an open problem belonging to ontology and natural language processing. Considering a word in a sentence, WSD identifies which of its senses is used in that sentence (for multi-sense words) [29].

6. Borin, L., Forsberg, M.: From the people's synonym dictionary to fuzzy synsets-first steps. In: Proc. LREC 2010 workshop Semantic relations. Theory and Applications. (2010).
7. Carstairs-McCarthy, A.: The origins of complex language : an inquiry into the evolutionary beginnings of sentences syllables and truth. (1999).
8. Dubois, D. et al.: On possibility/probability transformations. Proc. Fourth IFSA Conf. 103–112 (1993).
9. Dubois, D., Prade, H.: Unfair coins and necessity measures: Towards a possibilistic interpretation of histograms. Fuzzy Sets Syst. 10, 1-3, 15–20 (1983).
10. El-gedawy, M.N.: Using Fuzzifiers to Solve Word Sense Ambiguation in Arabic Language. Citeseer. 79, 2, 1–8 (2013).
11. Erk, K. et al.: Measuring word meaning in context. Comput. Linguist. 39, 3, 511–554 (2013).
12. Erk, K., McCarthy, D.: Graded word sense assignment. In: EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. pp. 440–449 (2009).
13. Esuli, A., Sebastiani, F.: SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation. pp. 417–422 (2006).
14. Fellbaum, C.: WordNet: An Electronic Lexical Database. Br. J. Hosp. Med. London Engl. 2005. 71, 3, 423 (1998).
15. Gonçalo Oliveira, H., Gomes, P.: Automatic Discovery of Fuzzy Synsets from Dictionary Definitions. In: 22nd Int. Joint Conf. on Artificial Intelligence. pp. 1801–1806 (2011).
16. Hirst, G.: Ontology and the Lexicon. In: Staab, S. and Studer, R. (eds.) Ontology and the Lexicon. pp. 269–292 Springer Berlin Heidelberg (2009).
17. Hurford, J.: Why Synonymy is Rare: Fitness is in the Speaker, http://www.isrl.uiuc.edu/~amag/langev/paper/hurford03ECAL.html, (2003).
18. Kann, V., Rosell, M.: Free construction of a free Swedish dictionary of synonyms. In: Proc. 15th Nordic Conf. on Comp. Ling.–NODALIDA (5). pp. 1–6 (2005).
19. Kilgarriff, A.: "I don't believe in word senses." Comput. Hum. 31, 2, 25 (1997).
20. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Mining Text Data. pp. 415–463 (2012).
21. Medhat, W. et al.: Sentiment analysis algorithms and applications: A survey. Ain Shams Eng. J. 5, 4, 1093–1113 (2014).
22. Miller, G. a. et al.: Introduction to wordnet: An on-line lexical database. Int. J. Lexicogr. 3, 4, 235–244 (1990).
23. Miller, G. a.: WordNet: a lexical database for English. Commun. ACM. 38, 11, 39–41 (1995).
24. Miller, G.A.: Nouns in wordnet. In: Fellbaum, C. (ed.) WordNet: An electronic lexical database. pp. 23–46 MIT press (1998).
25. Osgood, C.E.: the Nature and Measurement of Meaning. Psychol. Bull. 49, 3, 227 (1952).
26. Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Syst. 89, 14–46 (2015).
27. Rosso, P. et al.: Two Web-based approaches for noun sense disambiguation. In: Proceedings of 6th International Conference, CICLing 2005. pp. 267–279 Springer (2005).
28. Velldal, E.: A fuzzy clustering approach to word sense discrimination. In: Proceedings of the 7th International conference on Terminology and Knowledge Engineering. (2005).
29. Weaver, W.: Translation. Mach. Transl. Lang. 14, 15–23 (1955).
30. Zadeh, L. a.: Fuzzy Sets as a basis for a theory of possibility. Fuzzy Sets Syst. 1, 1, 3–28 (1978).