

A Vague Sense Classifier for Detecting Vague Definitions in Ontologies

Panos Alexopoulos

iSOCO S.A.

Madrid, Spain

palexopoulos@isoco.com

John Pavlopoulos

Department of Informatics,

Athens University of Economics and Business

Athens, Greece

annis@aueb.gr

Abstract

Vagueness is a common human knowledge and linguistic phenomenon, typically manifested by predicates that lack clear applicability conditions and boundaries such as *High*, *Expert* or *Bad*. In the context of ontologies and semantic data, the usage of such predicates within ontology element definitions (classes, relations etc.) can hamper the latter's quality, primarily in terms of shareability and meaning explicitness. With that in mind, we present in this paper a vague word sense classifier that may help both ontology creators and consumers to automatically detect vague ontology definitions and, thus, assess their quality better.

1 Introduction

Vagueness is a common human knowledge and language phenomenon, typically manifested by terms and concepts like *High*, *Expert*, *Bad*, *Near* etc., and related to our inability to precisely determine the extensions of such concepts in certain domains and contexts. That is because vague concepts have typically blurred boundaries which do not allow for a sharp distinction between the entities that fall within their extension and those that do not (Hyde, 2008) (Shapiro, 2006). For example, some people are borderline tall: not clearly “tall” and not clearly “not tall”.

Ontologies, in turn, are formal shareable conceptualizations of domains, describing the meaning of domain aspects in a common, machine-processable form by means of concepts and their interrelations (Chandrasekaran et al., January February 1999). As such, they are widely used for the production and sharing of structured data and knowledge that can be commonly understood among human and software agents.

When building ontologies and semantic data, engineers and domain experts often use predicates that are vague. While this is not always an intentional act, the use of such predicates influences in a negative way the comprehension of this data by other parties and limits their value as a reusable source of knowledge (Alexopoulos et al., 2013). The reason is the subjective interpretation of vague definitions that can cause **disagreements** among the people who develop, maintain or use a vague ontology. In fact, as shown in (Alexopoulos et al., 2013), vagueness in ontologies can be a source of problems in scenarios involving i) structuring data with a vague ontology (where disagreements among experts on the validity of vague statements may occur), ii) utilizing vague facts in ontology-based systems (where reasoning results might not meet users' expectations) and iii) integrating vague semantic information (where the merging of particular vague elements can lead to data that will not be valid for all its users).

In this context, our goal in this paper is to enable ontology producers (engineers and domain experts) as well as consumers (i.e., practitioners who want to reuse ontologies and semantic data) to detect, in an automatic way, ontology element definitions that are potentially vague. Such a detection will help ontology creators build more comprehensible and shareable ontologies (by refining, eliminating or just documenting vague definitions) and consumers assess, in an easier way, their usability and quality before deciding to use it.

Our approach towards such a detection involves training a classifier that may distinguish between vague and non-vague term word senses and using it to determine whether a given ontology element definition is vague or not. For example, the definition of the ontology class “*StrategicClient*” as “*A client that has a high value for the company*” is (and should be) characterized as vague while the definition of “*AmericanCompany*” as “*A com-*

pany that has legal status in the Unites States” is not. The classifier is trained in a supervised way, using vague and non-vague sense examples, carefully constructed from *WordNet*.

The structure of the rest of the paper is as follows. In the next section we briefly present related work while in section 3 we describe in detail our vague sense classifier, including the training data we used and the evaluation we performed. Section 4 describes the results of applying the classifier in an a publicly available ontology, illustrating its usefulness as an ontology evaluation tool. Finally, section 5 summarizes our work and outlines its future directions.

2 Related Work

The phenomenon of vagueness in human language and knowledge has been studied from a logic and philosophical point of view in a number of works (Hyde, 2008) (Shapiro, 2006) and different theories and paradigms have been proposed to accommodate it, including supervaluationism (Keefe, 2008), many-valued logic and fuzzy logic (Klir and Yuan, 1995). Moreover, in the context of ontologies, one may find several works focusing on acquisition, conceptualization and representation of vague knowledge, mainly following a fuzzy logic based approach (Bobillo and Straccia, 2011) (Stoilos et al., 2008) (Abulaish, 2009). Nevertheless all these approaches rely on manual identification and analysis of vague terms and concepts by domain experts and, to the best of our knowledge, no work attempts to automate this task.

Another set of related work consists of approaches for subjectivity and polarity labeling of word senses (Wiebe and Riloff, 2005) (Wiebe and Mihalcea, 2006) (Wilson et al., 2005) (Su and Markert, 2008) (Esuli and Sebastiani, 2006) (Akkaya et al., 2011). While vagueness is related to both phenomena (as polarized words are often vague and vague words are typically subjective), it is not exactly the same as these (e.g., subjective statements do not always involve vagueness) and, thus, requires specialized treatment. To illustrate that, we compare in subsequent sections our vague sense classifier with the subjective sense classifier of (Wilson et al., 2005), showing that the former performs significantly better than the latter.

3 Supervised Classification for Vague Term Detection

3.1 Data

We created a dataset of 2,000 adjective senses, collected from *WordNet*, such that 1,000 of them had a vague definition and the the rest a non vague definition. A sample of these senses is shown in Table 1 while the whole dataset, which to the best of our knowledge is the first of its kind, is publicly available for further research¹.

The dataset was constructed by an ontology expert. As the task of classifying a text as vague or not can be quite subjective, we asked from two other human judges to annotate a subset of the dataset’s definitions (100), and we measured inter-annotator agreement between all three. We found mean pairwise *JPA* (Joint Probability of Agreement) equal to 0.81 and mean pairwise *K* (Cohen’s Kappa) equal to 0.64, both of which indicate a reasonable agreement.

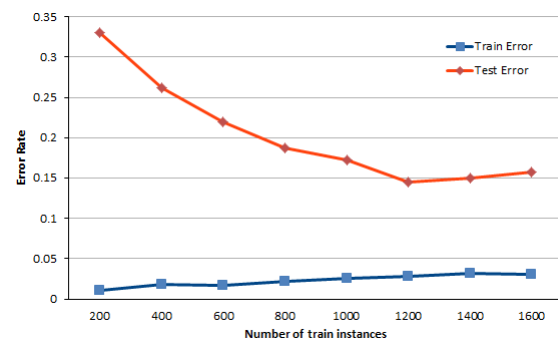


Figure 1: Train and test error rate, per number of training instances.

3.2 Training and Evaluation

We used the first 80% of the data (i.e., 800 vague and 800 non vague instances) to train a multinomial Naive Bayes classifier.² We removed stop words and we used the bag of words assumption to represent each instance.³ The remaining 20% of the data (i.e., 200 vague and 200 non vague instances) was used as a test set. Accuracy was found to be 84%, which is considerably high. In Figure 1, is shown the error rate on the test and train data, as we increase the number of training instances. We see that the two curves, initially,

¹<http://glocal.isoco.net/datasets/VagueSynsets.zip>

²We used the implementation of Scikit-Learn found at <http://scikit-learn.org/stable/>.

³We used the list of stopwords provided by Scikit-Learn.

Vague Adjectives	Non Vague Adjectives
Abnormal: not normal, not typical or usual or regular or conforming to a norm	Compound: composed of more than one part
Impenitent: impervious to moral persuasion	Biweekly: occurring every two weeks
Notorious: known widely and usually unfavorably	Irregular: falling below the manufacturer's standard
Aroused: emotionally aroused	Outermost: situated at the farthest possible point from a center
Yellowish: of the color intermediate between green and orange in the color spectrum, of something resembling the color of an egg yolk	Unfeathered: having no feathers

Table 1: Sample Vague and Non-Vague Adjective Senses

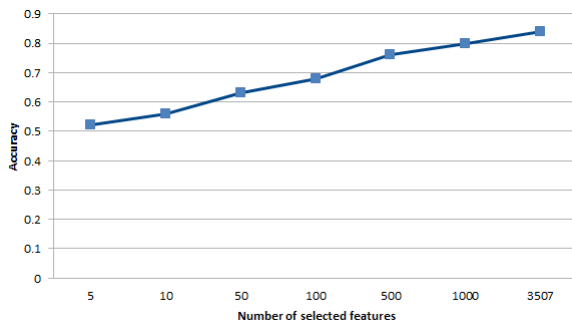


Figure 2: Accuracy on the test data, per number of selected features.

have a big gap between them, but this is progressively reduced. However, more (or more complicated) features could be beneficial; we intend to study this further in the future.

We also examined the hypothesis of the existence of a small set of words that are often found in vague definitions, but not in definitions which are not vague, as then it would be very easy for a system to use these words and discriminate between the two classes. To do this, we performed feature selection with the chi-squared statistic for various number of features and computed the accuracy (i.e., one minus the error rate). As we show in Figure 2, accuracy for only 5 selected features is 50%, which is the same as if we selected class in random. However, by increasing the number of selected features, accuracy increases significantly. This shows that there is not a subset of words which could be used to discriminate between the two classes; by contrast, most of the words play their role. Again, this is something to be further studied in future research.

Finally, in order to verify our intuition that

vagueness is not the same phenomenon as subjectiveness (as we suggested in section 2), we used the subjective sense classifier of (Wilson et al., 2005) to classify the data of section 3.1 as subjective or objective, assuming that vague senses are subjective while non-vague ones objective. The particular classifier is part of the OpinionFinder⁴ system and the results of its application in the 2000 adjective senses of our dataset were as follows. From the 1000 vague senses, only 167 were classified as subjective while from the 1000 non-vague ones 993. These numbers do not reflect of course the quality of OpinionFinder as a subjectivity detection system, they merely illustrate the fact that treating vagueness in the same way as subjectiveness is not really effective and, thus, more dedicated, vagueness-specific work is needed.

4 Use Case: Detecting Vagueness in CiTO Ontology

To evaluate the effectiveness and potential of our classifier for detecting vague ontological definitions, we considered a publicly available ontology called CiTO⁵. CiTO is an ontology that enables characterization of the nature or type of citations and consists primarily of relations, many of which are vague (e.g. the relation *cito:plagiarizes*). In order to compare the experts' vague/non-vague classification with the output of our system, we worked as follows. We selected 44 relations from CiTO (making sure to avoid duplications by e.g. avoiding having both a relation and its inverse) and we had again 3 human judges manually classify them as vague or not. In the end we got 27 vague

⁴<http://mpqa.cs.pitt.edu/opinionfinder/>

⁵<http://purl.org/spar/cito/>

Vague Relations	Non Vague Relations
<i>plagiarizes</i> : A property indicating that the author of the citing entity plagiarizes the cited entity, by including textual or other elements from the cited entity without formal acknowledgement of their source.	<i>sharesAuthorInstitutionWith</i> : Each entity has at least one author that shares a common institutional affiliation with an author of the other entity.
<i>citesAsAuthority</i> : The citing entity cites the cited entity as one that provides an authoritative description or definition of the subject under discussion.	<i>providesDataFor</i> : The cited entity presents data that are used in work described in the citing entity.
<i>speculatesOn</i> : The citing entity speculates on something within or related to the cited entity, without firm evidence.	<i>retracts</i> : The citing entity constitutes a formal retraction of the cited entity.
<i>supports</i> : The citing entity provides intellectual or factual support for statements, ideas or conclusions presented in the cited entity.	<i>includesExcerptFrom</i> : The citing entity includes one or more excerpts from the cited entity.
<i>refutes</i> : The citing entity refutes statements, ideas or conclusions presented in the cited entity.	<i>citesAsSourceDocument</i> : The citing entity cites the cited entity as being the entity from which the citing entity is derived, or about which the citing entity contains metadata.

Table 2: Sample Vague and Non-Vague Relations in CiTO

relations and 17 non-vague, a sample of which is shown in Table 2.

Then we applied the trained vagueness classifier of the previous section on the textual definitions of the relations. The results of this were highly encouraging; 36/44 (82%) relations were correctly classified as vague/non-vague with 74% accuracy for vague relations and 94% for non-vague ones. Again, for completeness, we classified the same relations with OpinionFinder (as in the previous section), in order to check if subjectivity classification is applicable for vagueness. The results of this were consistent to the ones reported in the previous section with the Wordnet data: 18/44 (40%) overall correctly classified relations with 94% accuracy for non-vague relations but only 7% for vague ones.

5 Conclusions and Future Work

In this paper we considered the problem of automatically detecting vague definitions in ontologies and we developed a vague word sense classifier using training data from *Wordnet*. Experiments with both *Wordnet* word senses and real ontology definitions, showed a considerably high accuracy of our system, thus verifying our intuition that vague and non-vague senses can be separable. We

do understand that vagueness is a quite complex phenomenon and the approach we have followed in this paper rather simple. Yet, exactly because of its simplicity, we believe that it can be a very good baseline for further research in this particular area. The vague/non-vague sense dataset we provide will be also very useful for that purpose.

Our future work comprises two main directions. On the one hand, as we mentioned in the introduction, we intend to incorporate the current classifier into an ontology analysis tool that will help ontology engineers and users detect vague definitions in ontologies and thus assess their quality better. On the other hand, we want to further study the phenomenon of vagueness as manifested in textual information, improve our classifier and see whether it is possible to build a vague sense lexicon, similar to lexicons that have already been built for subjectivity and sentiment analysis.

Acknowledgments

The research leading to this results has received funding from the People Programme (Marie Curie Actions) of the European Union’s 7th Framework Programme P7/2007-2013 under REA grant agreement *n*^o 286348.

References

- M. Abulaish. 2009. An ontology enhancement framework to accommodate imprecise concepts and relations. *Journal of Emerging Technologies in Web Intelligence*, 1(1).
- C. Akkaya, J. Wiebe, A. Conrad, and R. Mihalcea. 2011. Improving the impact of subjectivity word sense disambiguation on contextual opinion analysis. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL '11*, pages 87–96, Stroudsburg, PA, USA. Association for Computational Linguistics.
- P. Alexopoulos, B. Villazon-Terrazas, and Pan J.Z. Pan. 2013. Towards vagueness-aware semantic data. In Fernando Bobillo, Rommel N. Carvalho, Paulo Cesar G. da Costa, Claudia d'Amato, Nicola Fanizzi, Kathryn B. Laskey, Kenneth J. Laskey, Thomas Lukasiewicz, Trevor Martin, Matthias Nickles, and Michael Pool, editors, *URSW*, volume 1073 of *CEUR Workshop Proceedings*, pages 40–45. CEUR-WS.org.
- F. Bobillo and U. Straccia. 2011. Fuzzy ontology representation using owl 2. *International Journal of Approximate Reasoning*, 52(7):1073–1094, October.
- B. Chandrasekaran, J. Josephson, and R. Benjamins. January - February 1999. What are ontologies and why do we need them? *IEEE Intelligent Systems*, 14(1):Page 20–26.
- A. Esuli and F. Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06)*, pages 417–422.
- D. Hyde. 2008. *Vagueness, Logic and Ontology*. Ashgate New Critical Thinking in Philosophy.
- R. Keefe. 2008. Vagueness: Supervaluationism. *Philosophy Compass*, 3:315–324.
- G. Klir and B. Yuan. 1995. *Fuzzy Sets and Fuzzy Logic, Theory and Applications*. Prentice Hall.
- S. Shapiro. 2006. *Vagueness in Context*. Oxford University Press.
- G. Stoilos, G. Stamou, J.Z. Pan, N. Simou, and V. Tzouvaras. 2008. Reasoning with the Fuzzy Description Logic f-SHIN: Theory, Practice and Applications. pages 262–281.
- F. Su and K. Markert. 2008. From words to senses: A case study of subjectivity recognition. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1, COLING '08*, pages 825–832, Stroudsburg, PA, USA. Association for Computational Linguistics.
- J. Wiebe and R. Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of COLING-ACL 2006*.
- J. Wiebe and E. Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *In CICLing2005*, pages 486–497.
- T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. 2005. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations, HLT-Demo '05*, pages 34–35, Stroudsburg, PA, USA. Association for Computational Linguistics.