Automatically Clustering Ontological Annotated Sentences to Detect Semantic Frames

Alexandra Moreira, Alcione Oliveira de Paiva, and Giorgio Torres

Departamento de Informática, Universidade Federal de Viçosa (UFV),
CEP 36570-000, Viçosa MG,
Brazil

xandramoreira@yahoo.com.br,
{alcione,torres.giorgio}@gmail.com
http://www.dpi.ufv.br

Abstract. Lexical databases of semantic frames have been shown to be useful in problems related to natural language processing. However, creation of such databases is a task that is time consuming and involves many manual steps. One of these steps is selection and grouping of sentences to identify frames. However, we advocate that if sentences were previously annotated with ontological information, this grouping could be executed automatically. In this article we present tests performed with clustering sentences containing the lexeme Travel (noun and verb). Tests showed that the use of clustering algorithms on ontologically annotated sentences is a promising step towards automating construction of semantic frames databases.

Keywords: Clustering sentences, ontological annotation, frame semantics, FrameNet.

1 Introduction

The frame semantics proposed by Charles Fillmore [6] is a theory which states that the meaning of a lexeme can only be known from the knowledge of the scene where it occurs. Based on this theory, lexical databases, called FrameNet, describing the predicate-argument structure elements in a given scene were developed [20]. Lexical databases of semantic frames have been shown to be useful in problems related to natural language processing [4] [8] [13]. However, creation of such databases is a task that is time consuming and involves many manual steps [20]. One of these steps is the selection and grouping of sentences to identify frames. According to [20], The core of the process is to search for corpus attestations of a group of words that the FrameNet developers believe to have some semantic overlap. After that step they divide these attestations into groups and afterwards, combine the small groups into large enough groupings to make reasonable frames at which point we may (equivalently) call the words targets, lexical units, or frame-evoking elements. As one can see, the process is essentially manual, even with some auxiliary computational tools. However, automating this task is not a trivial process, since it requires a lot of common sense knowledge. We propose here to move up a step on the path to automate this process. We advocate that if sentences were previously annotated with ontological information, this grouping could be executed automatically. In this article we present tests performed with clustering sentences containing the lexeme Travel (noun and verb).

This article have the following structure: the next section presents the related work; section three succinctly presents the FrameNet; our proposal is presented in section four; section five presents the results and finally, section six presents the conclusions.

2 Related Works

Using semantic information to group or extracting information has been a subject widely investigated, nevertheless, no work that exploits corpus annotated with ontological types to perform groupings of sentences have been found. In [5] was presented a cooperative Machine Learning system which is able to acquire subcategorization verb frames with restrictions of selection and ontologies for specific domains from syntactically parsed technical texts in natural language. Texts and parsing may be noisy. The difference of this work is that the former extracts ontology instead of using it to detect the frame. Chow et al. [3] carried out a mapping between word-meanings (WordNet), frame-semantics (FrameNet) and world concepts captured by SUMO Ontology. The mapping provided a knowledge base for Semantic Role Labeling(SRL), identifying the appropriate range of possible semantic roles with respect to the event evoked by verb. In [1] was presented a research in Word Sense Disambiguation problem based on grouping noun representations of the senses. The proposal was based on the clustering of noun sense representations. In [10] is proposed an approach which utilizes ontology knowledge to automatically denote the implicit semantics of textual requirements. The authors state that "requirements documents include the syntax of natural language but not the semantics". They performed a semantic annotation of the requirements specification automatically and after this step is generated a domain model of the intended system. The common point with our work is the use of ontological annotation for analysis of sentences in natural language, however the scope and purpose differ widely from the present work.

3 The FrameNet

Frame Semantics arose as a response to the inability of traditional semantic to give account for different interpretations of lexical elements, such as explaining why it is not appropriate to characterize the Pope as a bachelor [9]. This is a classic example, used in several attestations [11] [16] [6] of the failure of the compositional semantic approach that defines a concept through minimum and necessary conditions. In fact, to understand the concept evoked by the lexical unit bachelor, one need to understand a chain of interrelated conceptual structures, such as the institution of marriage in western world, the notion of the typical functions of a married man and when one person is able to exercise those functions. Only then is possible to properly apply the term “bachelor” to someone. This is true for the majority of lexemes in natural language. Lexemes whose meaning can only be understood by understanding the entire concepts involved (gestalt) and not by their individual analysis.

FrameNet is a lexical semantic database based on Semantic Frames and supported by evidence from corpora. The pioneer FrameNet was developed by the International Computer Science Institute in Berkeley under the leadership of Collin F. Baker, Charles J. Fillmore and John B. Lowe [2]. The project aims to record the semantic and syntactic combinatorial possibilities (valences) of each predicative word (names, adjectives and verbs) in each of its senses. The basic concepts underlying the FrameNet project are the concepts of *frames*, *relations* between frames, *lexical units* (LU) and *frame elements* (FE). A lexical unit (LU) is the pairing of a word with a meaning [20]. According to the same author, each sense of a polysemous word belongs to a different semantic frame. A LU evokes a frame. For example, the occurrence of the word *buy* in a sentence invokes the event of a commercial purchase captured by the **Commerce_buy** frame. Frame Elements (FE) are roles that occur in a given frame. For example, the frame **Commerce_buy** describes common situations involving roles such as buyer, goods, seller, location and money. By presenting a particular frame, the system displays a definition and a list of elements of frames, and for each FE is presented a set of annotated sentences, extracted from a *corpus*. Frames are interconnected, forming a system of frames. They are connected through semantic relationships, such as *inheritance*, *use*, *subframe* and *perspective*. This differentiates them from other lexical databases, such as the thesauri. Semantic relations are asymmetrical frames forming a directed graph.

As already mentioned lexical databases such as a FrameNet are useful in a variety of natural language processing applications. However, the construction of a FrameNet is essentially a manual work with the support of some computational tools. The proposal described below seeks to contribute to increase the degree of automation of the process.

4 The Proposal

Ontological information imposes contextual constraints and help establish the scene that is taking place. Sentences belonging to the same scene will contain the same ontological types or ontological types closely related. Adding of an annotation step to the FrameNet development process to add ontological type information is advocated by [15]. However the addition of such information is not a trivial task. There are some projects that address the task of ontological annotation, such as [17] and [21]. Here we report an use of this annotation layer with the objective of helping the grouping of sentences for extracting semantic frames. Automatic annotation of ontological information is also being addressed within this project but there are still no published results. Fig. 1 summarizes the steps of the clustering process.

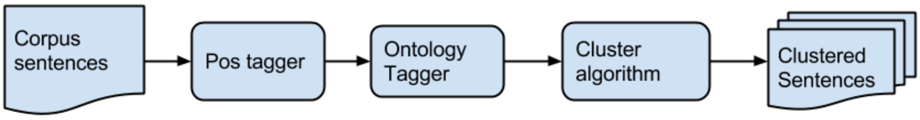


Fig. 1. The Clustering Process.

Part of speech (POS) annotation stage is important to help the ontology annotation step. After its lexical class had been identified is easier to identify the type of a term. To perform clustering the framework Weka was used. Weka [7] is a set of programs written in the Java programming language and is oriented carry out data mining and machine learning tasks. The Weka was developed by the University of Waikato in New Zealand and is an open source tool. The tool can be run directly or incorporated into other programs and provides tools for pre-processing, classification, regression, clustering and data visualization. The Framework has several clustering algorithms, which allows performing many tests within the same environment.

5 Results

To test the proposed system it was used a *subcorpus* of the *Corpus do Português* available for free access in the BRIGHAM YOUNG UNIVERSITY portal [1]. The subcorpus consists of sentences containing the lexeme “Travel”, both the verb and the nominal in Brazilian Portuguese language. This subcorpus was used by [14] in the characterization of frame TRAVEL. In [9] the sentences were manually classified into *prototypical*, *quasi* and *metaphorical.* The prototypical class groups the typical sentences of the central meaning of the lexeme *travel.* That is: a displacement event to a particular locality executed by a conscious entity or group of entities, by themselves or by a transport means and for some purpose [2]. The *quasi* class groups the sentences that deviate in varying degrees from this central sense. The metaphorical class groups the sentences where the lexeme *travel* occurs in a metaphorical sense (e.g., time travel, spiritual, etc.). This is a good *corpus* to test whether the system will group in the same way sentences were manually grouped. 57 sentences were used as input to the system. 15 of these 57 sentences were previously classified as prototypical, 5 were classified as metaphorical, and 37 were previously classified as *quasi.*

[1] `http://corpus.byu.edu`

[2] `https://framenet.icsi.berkeley.edu/fndrupal/`

The ontology used was the SIMPLE-CLIPS ontology, (*Semantic Information for Multifunctional Plurilingual Lexica-Corpora e Lessici dell'Italiano Parlato e Scritto*) [12]. The SIMPLE-CLIPS ontology is based on *qualia* structure [18] and consists of semantic types organized through hierarchical and non-hierarchical conceptual relations. *Qualia* structure describes the nature of denotation through their fundamental attributes organized in formal, constitutive, telic and agentive dimensions. Ontological annotation was performed semi-automatically in [14].

Lexical items were annotated with the following semantic types: *human*, *vehicle*, *animal*, *abstract* and *local*. Occurrence or absence of these elements were used to create the vector space used by the clustering algorithm. Table 5 shows the attributes present in each sentence. The last attribute indicates the classification assigned by the human expert.

Table 1. Attributes present in each sentence.

vehicle local prototypical	vehicle local prototypical	vehicle abstract metaphorical
null quasi	vehicle local prototypical	human quasi
local prototypical	null quasi	human local prototypical
local prototypical	null quasi	animal quasi
human local prototypical	local prototypical	local quasi
vehicle prototypical	vehicle quasi	null quasi
human local prototypical	null quasi	human quasi
human vehicle prototypical	human quasi	human local prototypical
local prototypical	vehicle quasi	null quasi
null quasi	vehicle local prototypical	local quasi
null quasi	null quasi	vehicle quasi
animal abstract metaphorical	local quasi	animal quasi
local quasi	null quasi	null quasi
local quasi	animal quasi	vehicle local quasi
human quasi	abstract metaphorical	human quasi
human vehicle local prototypical	abstract quasi	null quasi
human quasi	local quasi	abstract metaphorical
human quasi	human quasi	null quasi
local quasi	abstract metaphorical	vehicle quasi

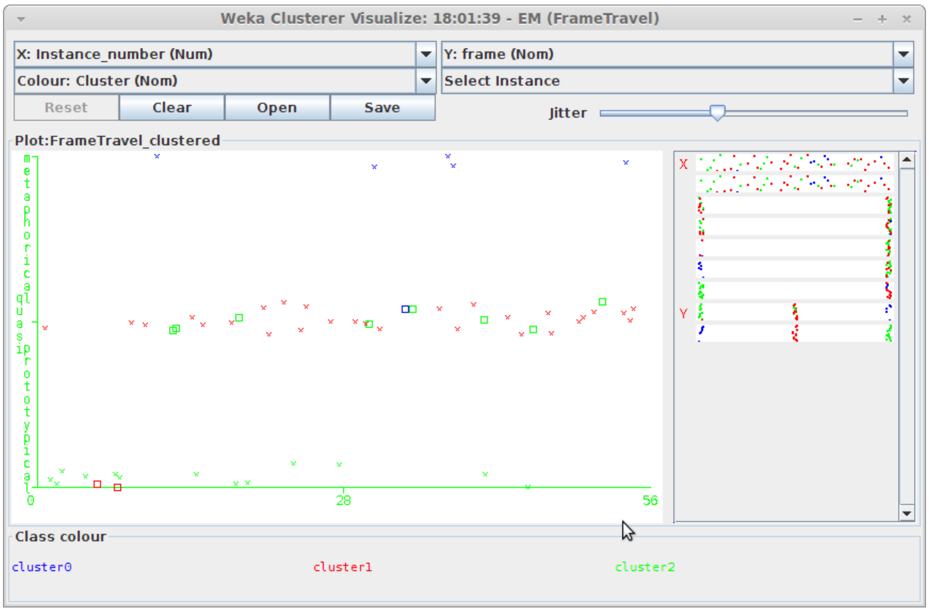
Those attributes lists were used as input for classification algorithms of Weka Framework. EM (expectation maximisation) algorithm was the one with best results. EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters [7]. EM can decide how many clusters to create by cross validation, or one may specify *apriori* how many clusters to generate. It disagreed with the classification done by humans in 19%. This rate seems high at first glance, however, it is necessary to analyze this result more carefully. Fig. 2 presents part of the textual output of the Simple EM algorithm and Fig. 3 shows the plot of the Clustering.

6 Conclusion

Tests showed that the use of clustering algorithms on ontologically annotated sentences is a promising step towards automating the construction of semantic

```
=== Run information ===

Scheme:weka.clusterers.EM -I 100 -N -1 -M 1.0E-6 -S 100
Relation:     FrameTravel
Instances:    57
Attributes:   7
              human
              vehicle
              animal
              abstract
              local
Ignored:
              num
              frame
Test mode:Classes to clusters evaluation on training data

=== Model and evaluation on training set ===

Clustered Instances

0       6 ( 11%)
1      30 ( 53%)
2      21 ( 37%)

Log likelihood: -2.28633

Class attribute: frame
Classes to Clusters:

  0  1  2  <-- assigned to cluster
  0  2 13 | prototypical
  1 28  8 | quasi
  5  0  0 | metaphorical

Cluster 0 <-- metaphorical
Cluster 1 <-- quasi
Cluster 2 <-- prototypical

Incorrectly clustered instances :    11.0     19.2982 %
```

Fig. 2. Part of the textual output of the algorithm (Simple EM - expectation maximisation).



Weka Clusterer Visualize: 18:01:39 - EM (FrameTravel)

X: Instance_number (Num)

Y: frame (Nom)

Colour: Cluster (Nom)

Select Instance

Plot:FrameTravel_clustered

Fig. 3. Plot of the clustering (Simple EM - expectation maximisation).

frames databases. In typically sentences related to the frame it can be noted a reasonable degree of accuracy in two classical clustering algorithms. The disagreements are most common in sentences with few annotation or difficult to be framed even by people. The use of more accurate ontological types annotation algorithms should lead to better results. As future work, we are analyzing the semantic annotator Wmatrix [19] to enable a broader analysis of larger *corpus*.

Acknowledgments. This research is supported in part by the funding agencies FAPEMIG, CNPq and by the Gapso company.

References

1. Anaya-Sánchez, H., Pons-Porrata, A., Berlanga-Llavori, R.: Word sense disambiguation based on word sense clustering. In: Advances in Artificial Intelligence-IBERAMIA-SBIA 2006, pp. 472–481. Springer (2006)
2. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1. pp. 86–90. Association for Computational Linguistics (1998)
3. Chow, I.C., Webster, J.J.: Mapping framenet and sumo with wordnet verb: Statistical distribution of lexical-ontological realization. In: Artificial Intelligence, 2006. MICAI'06. Fifth Mexican International Conference on. pp. 262–268. IEEE (2006)

4. Dannélls, D.: Applying semantic frame theory to automate natural language template generation from ontology statements. In: Proceedings of the 6th International Natural Language Generation Conference. pp. 179–183. Association for Computational Linguistics (2010)
5. Faure, D., Nédellec, C.: A corpus-based conceptual clustering method for verb frames and ontology acquisition. In: LREC workshop on adapting lexical and corpus resources to sublanguages and applications. vol. 707, p. 30 (1998)
6. Fillmore, C.J.: Scenes-and-frames semantics. Linguistic structures processing 59, 55–88 (1977)
7. Holmes, G., Donkin, A., Witten, I.H.: Weka: A machine learning workbench. In: Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on. pp. 357–361. IEEE (1994)
8. Johansson, R., Nugues, P.: A framenet-based semantic role labeler for swedish. In: Proceedings of the COLING/ACL on Main conference poster sessions. pp. 436–443. Association for Computational Linguistics (2006)
9. Katz, J.J., Fodor, J.A.: The structure of a semantic theory. language pp. 170–210 (1963)
10. Körner, S.J., Landhäußer, M.: Semantic enriching of natural language texts with automatic thematic role annotation. In: Natural Language Processing and Information Systems, pp. 92–99. Springer (2010)
11. Lakoff, G.: The invariance hypothesis: Is abstract reason based on image-schemas? Cognitive Linguistics (includes Cognitive Linguistic Bibliography) 1(1), 39–74 (1990)
12. Lenci, A., Busa, F., Ruimy, N., Gola, E., Monachini, M., Calzolari, N., Zampolli, A., Pustejovsky, J., Ogonowski, A., McCawley, C., et al.: Simple linguistic specifications. Deliverable D2 1 (2000)
13. Lenci, A., Montemagni, S., Venturi, G., Cutrulla, M.G.: Enriching the isst-tanl corpus with semantic frames. In: LREC. pp. 3719–3726 (2012)
14. Moreira, A., Salomão, M.M.M.: Applying bayesian networks and ontological types into lexeme to estimate the pertinence to a semantic frame. Revista Veredas 17(1), 149–164 (2013)
15. Moreira, A., Salomão, M.M.M.: Análise ontololólogica aplicada ao desenvolvimento de frames. ALFA: Revista de Linguística 56(2) (2012)
16. Petruck, M.R.: Frame semantics. Handbook of pragmatics pp. 1–13 (1996)
17. Pradhan, S.S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., Weischedel, R.: Ontonotes: A unified relational semantic representation. International Journal of Semantic Computing 1(04), 405–419 (2007)
18. Pustejovsky, J.: The generative lexicon. Computational linguistics 17(4), 409–441 (1991)
19. Rayson, P.: From key words to key semantic domains. International Journal of Corpus Linguistics 13(4), 519–549 (2008)
20. Ruppenhofer, J., Ellsworth, M., Petruck, M.R., Johnson, C.R., Scheffczyk, J.: FrameNet II: Extended theory and practice (2006)
21. Sanfilippo, A., Tratz, S., Gregory, M., Chappell, A., Whitney, P., Posse, C., Paulson, P., Baddeley, B., Hohimer, R., White, A.: Ontological annotation with wordnet. In: Proceedings of the International WordNet Conference GWC (2006)