Using Topic Salience and Connotational Drifts to Detect Candidates to Semantic Change

Armelle Boussidan
L2C2, Institut des Sciences Cognitives - CNRS, Université de Lyon, Bron, France
armelle.boussidan@isc.cnrs.fr

Sabine Ploux
L2C2, Institut des Sciences Cognitives - CNRS, Université de Lyon, Bron, France
sploux@isc.cnrs.fr

Abstract

Semantic change has mostly been studied by historical linguists and typically at the scale of centuries. Here we study semantic change at a finer-grained level, the decade, making use of recent newspaper corpora. We detect semantic change candidates by observing context shifts which can be triggered by topic salience or may be independent from it. To discriminate these phenomena with accuracy, we combine variation filters with a series of indices which enable building a coherent and flexible semantic change detection model. The indices include widely adaptable tools such as frequency counts, co-occurrence patterns and networks, ranks, as well as model-specific items such as a variability and cohesion measure and graphical representations. The research uses ACOM, a co-occurrence based geometrical model, which is an extension of the Semantic Atlas. Compared to other models of semantic representation, it allows for extremely detailed analysis and provides insight as to how connotational drift processes unfold.

1 Introduction

Semantic change has long been analyzed and theorized upon in historical linguistics. Its abstract and ungraspable nature made its detection a difficult task for computational semantics, despite the many tools available from various models of lexical treatment. Most extant theories are based on manual analysis of century long semantic drifts. From these works we inherit various typologies and repertories of causes of change (e.g., Bloomfield (1933)). However these types of analyses may not be suited to the large scale production of text in our societies. Not only has the quantity of produced text rocketed but its diffusion and speed of transmission has radically increased. In this context, recent studies have yielded promising results, showing that computational models of semantics can deal with assessed semantic change examples as well as detect candidates in corpora. Among them, some include topic salience as an index and others do not, as they rather try to quantify semantic change with reliable measures. In an era of information overflow, topic change takes on a new linguistic value, as it may be responsible for extremely quick paced semantic change, which can be ephemeral or become fixed. Topic salience might as well be a sociologically induced or press phenomenon with no semantic impact at all. However when both topic salience and connotational drift take place, a semantic phenomenon may be at stake. Our analysis is anchored in this process. We shall briefly introduce other approaches, explain our methods and the structure of our detection prototype (in progress) as well as give preliminary results before concluding with a discussion.

2 Measuring semantic change : previous work

To measure semantic change, one has to evaluate the semantics of a lexical item at a given point. To do so, semantic similarity measures in vector spaces or geometrical spaces may be used to compare the

item with its own occurrences at later points. This method has been applied in Sagi et al. (2009), where semantic *density* was calculated as the average angle between vectors in a semantic space. The *variability* of that density was observed for the same lexical item at different points in time. Density measures were applied to a series of acknowledged semantic change cases, in the *Project Gutenberg* Corpus, a historical corpus of English organized by documents. Results mostly include broadening and narrowing cases. The same method yielded results on the difference between nominal and verbal types of change, showing that verbs were more likely to change than nouns (Sagi (2010)).
Cook and Stevenson (2010) also used assessed cases from the historical linguistics literature. They detected changes in the semantic orientation of words (or polarity shifts) namely amelioration and pejoration. They then applied this methodology to detect possible un-assessed candidates. They used three English corpora as corpus slices, covering approximately a four century time-span.
Volatility has also been assessed by Holz and Teresniak (2010), who adapted a measure from econometrics to quantify semantic change in a time sliced corpus. The volatility measure relied on the computation of the rank series for every co-occurent term and on the coefficient of variation of all co-occurrent terms (Holz and Teresniak (2010)). The method was applied to search words in modern corpora in German and English (the *Wortschatz* and *the New York Times*). The strong point of this measure is that it is independent from word frequency, however it does not provide detailed analysis about the underlying semantic processes.

3 Methods

Of the three cited works, our approach is closer to that of Holz and Teresniak (2010) in that both their work and ours are conducted on very recent corpora. We are currently conducting short diachrony detection, analysis and representation on a modern press corpus in French (the newspapers *Le Monde*, 1997-2007). We use the ACOM model (Ji et al. (2003)) an extension of the Semantic Atlas Model (Ploux et al. (2010)) that uses factor analysis to provide geometrical representations of word co-occurrence in corpus (both models are freely available on `http://dico.isc.cnrs.fr/eng/index.html`). The model relies on *cliques*, which are organized subsets of co-occurrent words, from which clustering can be made. To extract co-occurrent words, we apply ACOM on a time-sliced corpus. For each slice t, a word-association table is constructed using all headwords (see Ploux et al. (2010) for a complete methodological description). Each headword $W_t^i (1 \leq i \leq N$, where N is the total number of types in the corpus slice) has children $(c_j s)$ that are arranged in descending order of co-occurrence with W_t^i [1]:

$$W_t^i : c1; c2; \ldots; cn$$

We apply two factors to filter this table: α where $0 \leq \alpha \leq 1$ to eliminate the rarely co-occurring children of W_n^i :

$$W_t^i : c1; c2; \ldots; ck$$

where $k = n\alpha$ and n is the original number of children of W_t^i, and β where $\beta (0 \leq \beta \leq 1)$ to cut off rarely co-occurring of children of c_j:

$$(c_j^m : g1; g2; \ldots; gl (1 \leq j \leq k; l = m\beta))$$

On the basis of that table, cliques are calculated. The notion of clique is taken from graph theory (on graph therory see for ex. Golumbic (2004)). Mathematically, cliques are maximum connected subgraphs. In our case, the nodes are contexonyms. Then, correspondence factor analysis is applied (Benzécri (1980)) and the χ^2 distance is calculated between pairs of cliques to obtain a multidimensional space. A hierarchical clustering algorithm clusters cliques in thematic sets at several degrees of detail. Clusters show broad topic shifts whereas the cliques show fine-grained sub-organisation. Therefore the model allows for very detailed analysis as well as topical analysis. It also provides a graphic visualization for the semantics of a word. With the time-sliced corpus, we may extract maps for each subpart of the

[1] Children with co-occurrences under a 10,000th of the global frequency of the headword W_t^i are removed to reduce noise.

corpus and compare the spaces generated for the same word at different points in time, to complete the analysis.

3.1 Structure of the detection prototype

Currently our model is structured as follows: the corpus is transformed into a time-sliced ACOM database, with word frequencies and co-occurence frequencies. We apply an adjustable standard deviation filter to extract significant frequency and co-occurrence frequency variations as well as co-occurrence network variations. (The co-occurrence window is adjustable to the sentence, paragraph or other window sizes). If we only detect frequency variation, there is a suspicion that the headword might undergo context variation later, but it could also be an ephemeral press or fashion phenomenon with no semantic impact. However if we detect both significant frequency variations and co-occurrence variations, there is a higher chance that the context variations are a reflection of semantic variation. At this stage we apply indices based on rank variation, clique analysis and clique-term variation analysis (described in Boussidan et al. (2010)) as well as manual analysis to determine the nature of the change. The next step to verify that the item has undergone semantic change is its stabilization over time. This detection path highlights short diachronic change. We may also detect significant co-occurence variations with no significant headword frequency variation, in which case we may apply directly the indices to check whether the context shifts reveal an anchored meaning shift. If the indices highlight a meaning shift, the former is necessarily much more subtle than the short diachronic change that we detected previously. It might be the reflection of a longer term process of which the trigger might not be contained in the given corpus.

4 Preliminary results

4.1 Testing examples

To conceive a detection model, we first conducted experiments using attested examples or using words that we selected after manually observing that a shift was taking place. By testing these examples, we could extract data about how the model would render them so as to use it to create detection indices and parameters. Among these was the French word *malbouffe* (literally "bad grub" or "junk food"), a neology selected from a previously established list of new dictionary entries (Martinez (2009)). The corpus showed how the different spellings of the words alternated before yielding the current one. Analysis of the co-occurrence networks showed that one of the most important co-occurrent words, *Bové*, the name of a French political actor, had almost the same co-occurrence network as *malbouffe*. From this observation and after comparing definitions and previous contexts of use, we could infer that this person gave the word *malbouffe* its new meaning, by superimposing political values on it, on top of its dietetic values. Co-occurrence networks therefore allowed us to analyse the process of meaning shift. The full analysis of this example may be found in Boussidan et al. (2009).

We also tested a more subtle connotational drift with the word *mondialisation* ("globalization"), which undergoes clear contextual change in the corpus. The word first appeared in contexts defined by the political, economical and intellectual positions it brings about, with strong co-occurrents such as *défi* (“challenge”), *progrès* (“progress") or *menace* (“threat”). It then drifted into a complete network of words related to one single French political movement of anti-globalization in 2001. Therefore the use of *mondialisation* gained a new connotation, whereas its synonym *globalisation* ("globalization") remained quite neutral politically. The analysis of this example revealed that some terms were used as pivots, providing linkage between the existing cliques and the new ones. Pivots therefore provided a good tool to observe meaning re-organisation. The full analysis of this example may be found in Boussidan et al. (2010) and the corresponding dynamic representation on `http://dico.isc.cnrs.fr/en/diachro.html`.

4.2 Semantic change detection

On the basis of these preliminary examples, we designed a semantic change detection prototype. Testing examples brought to light the difficulty of discriminating press-related topic salience with no

semantic impact from topic salience with a semantic impact. Detection is conducted in three stages. The first stage relies on frequency variation to extract topic variations of context in the corpus. For instance by setting the filter to retain words for which the coefficient of variation[2] is higher than 0.5, we obtain a list of words that may be classified into three loose semantic sets and a fourth set grouping all independent items. These semantic sets include words related to:

- war, terrorism and violence
- technology
- illness

By adjusting the settings we mayalso include more subtle topic variations if needed or conversely, looser ones. The second stage involves co-occurrence variation so as to extract the changes in semantic networks and thus in connation, for a lexical item. For instance, we detected that the word *logiciel* ("software") underwent a frequency co-occurrence peak with *libre* ("free") in January 2001. The expression *logiciel libre* stands for "freeware" and has been renamed *gratuiciel* or *graticiel* (a blending of *gratuit*, "free" with *logiciel*, "software") in Quebec. We therefore detect a new compositional expression that coins a French equivalent to the word *freeware* used until then.

Another example of connotational drift is the word *navigation* ("navigation") which is only attested in the TLFI[3] and the Dictionnaire Historique de la Langue Française (Rey et al. (1998)), under the meaning relating to transport, firstly on seas and rivers and then via plane or spaceship. However, between 1997 and 2001 the word takes on a new major meaning in internet search, meaning "browsing". This is aparent when looking at the co-occurrence patterns of *navigation* with words related to technology and comparing them with co-occurrences of words related to transport. The technology words show peaks between 1997 and 2001 and then lower frequencies until 2007, whereas the transport words show stable use all the way through the corpus. The new use of *navigation*, however is almost obsolete now in spoken speech -or at least it has gone out of fashion- but the semantics of *navigation* have clearly integrated an additional domain and broadened. A simple search of French results on Google provides 5,500,000 documents for *navigation internet*, among which are a lot of recent ones. However the meaning *to search the internet* grew from the name of a specific web navigator: the *Netscape Navigator* which was widespread in the 1990s but is no longer supported nowadays.

Both previous stages provide us with candidates to semantic change. The last stage is the stabilization of a connotational drift, whether it is a broadening, a narrowing, a domain shift or other. We are currently working on this last index. We often find that when a word undergoes semantic change, it goes through a phase of onomasiological competition in which other possible candidates may in turn become the new bearers of certain meanings. For *navigation* for instance, the word *surf* was a competitor, however both words now sound obsolete. It may be that none of them wins the competition, in which case the concept has become so deeply anchored in language and society that it does not need naming any more.

5 Discussion and Future Work

Since semantic phenomena, whether synchronic or diachronic, are very much corpus specific, it is difficult to conceive of a large scale universal detection method for them. However, tools may be built to be highly flexible in order to allow users to adjust settings to adapt to the corpus they deal with. This flexibility may encompass genre and stylistic variations when working with the same language as well as adaptation to a completely different language. We are considering global evaluations of the corpora's stylistics to avoid the detection of corpus specfic phenomena instead of broader language phenomena.

[2]The coefficient of variation is the ratio of the standard deviation to the mean

[3]`http://atilf.atilf.fr/tlf.htm`

Ideally the model should also be able to deal with timescale differences.The precise adjustment of these settings is part of our future research avenues along with incorporating an index for stabilization. This last filter is particularly difficult to create when dealing with ongoing phenomena. We may sometimes need to wait a few years to be able to establish whether a semantic change has stabilized.
To summarize, we are currently developing a filtering tool to extract candidates to semantic change on the basis of topic salience variation in corpus and co-occurrence network variation. Our approach shed light on the emergence of these phenomena at a very detailed level. Preliminary results showed that the tool was succesful at extracting those candidates; however it is not yet advanced enough to discriminate between context changes that affect a word without semantic impact and those that do have a semantic impact. This aspect constitutes our current research perspective.

6 Acknowledgements

This research is supported by the Région Rhône-Alpes, via the Cible Project 2009. Many thanks to Sylvain Lupone, previously engineer at the L2c2 for the tools he developed in this research's framework.

References

Benzécri, J.-P. (1980). *L'analyse des données : l'analyse des correspondances*. Paris: Bordas.

Bloomfield, L. (1933). *Language*. New York: Allen and Unwin.

Boussidan, A., S. Lupone, and S. Ploux (2009). La malbouffe : un cas de néologie et de glissement sémantique fulgurants. In *"Du thème au terme, émergence et lexicalisation des connaissances"*, Toulouse, France. 8 ème conférence internationale Terminologie et Intelligence Artificielle.

Boussidan, A., A.-L. Renon, C. Franco, S. Lupone, and S. Ploux (2010). Vers une méthode de visualisation graphique de la diachronie des néologies. Tubingen, Germany. Colloque Néologie sémantique et Corpus. in press.

Cook, P. and S. Stevenson (2010). Automatically identifying changes in the semantic orientation of words. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, Valletta,Malta. LREC 2010.

Golumbic, M. (2004). *Algorithmic graph theory and perfect graphs*. North-Holland.

Holz, F. and S. Teresniak (2010). Towards automatic detection and tracking of topic change. *Computational Linguistics and Intelligent Text Processing*, 327–339.

Ji, H., S. Ploux, and E. Wehrli (2003). Lexical knowledge representation with contexonyms. *Proceedings of the 9th Machine Translation Summit*, 194–201.

Martinez, C. (2009). *L'évolution de l'orthographe dans les Petit Larousse et les Petit Robert 1997-2008: une approche généalogique du texte lexicographique*. Ph. D. thesis, Université de Cergy-Pontoise.

Ploux, S., A. Boussidan, and H. Ji (2010). The semantic atlas: an interactive model of lexical representation. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, Valletta, Malta. LREC 2010.

Rey, A., T. Hordé, and L. Robert (1998). *Dictionnaire historique de la langue française : contenant les mots français en usage et quelques autres délaissés, avec leur origine proche et lointaine*. Paris.

Sagi, E. (2010). Nouns are more stable than verbs: Patterns of semantic change in 19th century english. Portland, OR. 32nd AnnualConference of the Cognitive Science Society. to be published.

Sagi, E., S. Kaufmann, and B. Clark (2009). Semantic density analysis: Comparing word meaning across time and phonetic space. In *GEMS: GEometrical Models of Natural Language Semantics*. EACL.