Neoveille, a Web Platform for Neologism Tracking

Emmanuel Cartier
Université Paris 13 Sorbonne Paris Cité - LIPN - RCLN UMR 7030 CNRS
99 boulevard Jean-Baptiste Clément
93430 Villetaneuse, FRANCE
emmanuel.cartier@lipn.univ-paris13.fr

Abstract

This paper details a software designed to track neologisms in seven languages through newspapers monitor corpora. The platform combines state-of-the-art processes to track linguistic changes and a web platform for linguists to create and manage their corpora, accept or reject automatically identified neologisms, describe linguistically the accepted neologisms and follow their lifecycle on the monitor corpora. In the following, after a short state-of-the-art in Neologism Retrieval, Analysis and Life-tracking, we describe the overall architecture of the system. The platform can be freely browsed at www.neoveille.org where detailed presentation is given. Access to the editing modules is available upon request.

1 Credits

Neoveille is an international Project funded by the ANR IDEX specific funding scheme. It gathers seven Linguistics and Research Centers. See website for details.

2 Introduction

Linguistic change is one of the fundamental properties of language, even if, at least on a short-term period, languages appear to be extremely conservative and reluctant to change. Whereas NLP efforts have mainly focused on synchronic language analysis, research and applications are very sparse on the diachronic side, especially concerning short term diachrony. But, with the availability of big web corpora, the maturity of automatic linguistic analysis and especially those able to process big data while maintaining a reasonable quality, it is now possible to monitor language change and track linguistic innovations.

3 Previous Work in Neology and Neology Tracking

Linguistic change has been studied for decades and even centuries in linguistics, and has been dealt with more recently in computational linguistics.

3.1 Linguistic Neology Models

3.1.1 Neologism Categories

Linguistic change has been first focused on by the Comparative Grammars School, whose main goal was to study languages diachronically. They have mainly based their descriptions and analysis on linguistic forms, describing phonetical, phonological, morphological, syntactical and semantic change on a long-term basis (Geeraerts, 2010). More recently, several attempts have emerged in the field of linguistic change, mainly focusing on the lexical units and proposing typology of neologisms (Schmid, 2015),(Sablayrolles, 2016).

3.1.2 Synchrony, Diachrony, Diastraty

A complementary approach is due to (Gevaudan and Koch, 2010) who state that every lexical evolution can be described through three parameters : two are universal, the semantic parameter (explicating a continuity of meaning or a discontinuity, in this case further described) and the stratic parameter (linking the linguistic structure to its sociological context : borrowings are explained this way); the third one is linked to every specific linguistic formal structure, with four generic matrices : conversion, morphological extension, composition and clipping).

Proceedings of the EACL 2017 Software Demonstrations, Valencia, Spain, April 3-7 2017, pages 95–98

3.1.3 Neologism Life-cycle(s)

Neology is one of the aspect of linguistic change, with necrology and stability. One important aspect is thus to model the lifecycle of a neologism, from its first occurrence to its potential disappearance or conventionalization. (Traugott and Trousdale, 2013) have proposed three salient states : innovation, propagation and conventionalisation, linking each to several more-or-less obvious properties. With these models in mind, NLP has developed several algorithms to track and study the lifecycle of neologisms.

3.2 Computational Models of Neology

Computational Linguistics has begun to work on linguistic change not long ago, mainly because it needs to have at hand large diachronic electronic corpora. Neology is moreover still considered as an secondary topic, as novel lexical units represents less than 5 percent of lexical units in corpora, according to several studies. But linguistic change is the complementary aspect of the synchronic structure. Every lexical unit is subjected to time, form and meaning can change, due to diastratic events and situations. The advent of electronic (long and short-term) diachronic corpora, scientific research and advances on word-formation and machine learning techniques able to manage big corpora, have permitted the emergence of neology tracking systems. Apart from a best knowledge of language lifecyle(s), these tools would permit to update lexicographic resources, computational assets and parsers.

From the CL point of view, the main questions are : how can we automatically track neologisms, categorize them and follow their evolution, from their first appearance to their conventionalisation or disappearance? At best, can we induce neology-formation procedures from examples and therefore predict potential neologisms?

3.3 Existing Neology Tracking System

3.3.1 the Exclusion Dictionary Architecture (EDA)

The main achievement of neology tracking consists in system extracting novel forms from monitor corpora, using lexicographic resources as a reference exclusion dictionary to induce unknown words, what we can call the ”exclusion dictionary architecture” (EDA). The first system is due to (Renouf, 1993) for English : a monitor corpora and a reference dictionary from which unknown words can be derived. Further filters then apply to eliminate spellings errors and Proper Nouns. Subsequent developments all replicate this architecture : OBNEO (Cabré and De Yzaguirre, 1995), NeoCrawler (Kerremans et al., 2012), Logoscope (Gérard et al., 2014) and more recently Neoveille (Cartier, 2016).

Four main difficulties arise from these architecture : first, EDA can not track semantic neologisms, as they use existing lexical units to convey innovative meanings; second, the design of a reference exclusion dictionary is not that obvious as it requires the existence of a machine-readable dictionary : this entails specific procedures to apply this architecture to less-resourced languages, and the availability of an up-to-date machine-readable dictionary for more resourced languages ; third, the EDA architecture is not sufficient in itself : among unknown forms, most of them are Proper Nouns, spelling mistakes and other cases derived from corpus boilerplate removal : this entails a post-processing phase to depart cases; Fourth, these systems do not take into account the sociological and diatopic aspects of neologism, as they limit their corpora to specific domains : a ideal system should be able to extend its monitoring to new corpora and maintain diastratic meta-datas to characterize novel forms. To the best of our knowledge, Neoveille (Cartier, 2016) is the only system implementing this aspect.

3.3.2 Semantic Neology Approaches

As for semantic neology, three approaches have been recently proposed, none of them being exploited in an operational system. The first one stems from the idea that meaning change is linked to domain change : every texts and thus the constituent existing lexical units are assigned one or more topic; if a lexical unit emerges in a new domain, a change in meaning should have occurred (Gérard et al., 2014). The main drawback of this approach is that it is limited to specific semantic change (it can not tackle conventional metaphors if appeared in the same domain, nor detect extension or restriction of meaning) and mainly limited to Nouns.

An other approach is linked to the distributional paradigm : ”You shall know a word by the company it keeps”(Firth, 1957). The main idea is to retrieve from a large corpora all the collocates or collostructions, and classify them according to sev-

eral metrics. The main salient resulting context words represent the "profile" (Blumenthal, 2009) or "sketch" (Kilgarriff et al., 2004) of a lexical unit for the given synchronic period. The most elaborated system is surely the Sketch Engine system, which propose for every lexical engine its "sketch", i.e. a list, for any user-defined syntactic schemas (for example modifiers, nominal subject, object and indirect object for verb) of occurrences, sorted by one or several association measure. This system can be improved in two main ways : first, it does not propose complete syntactic schemas for lexical units like verbs (it is limited to either a SUBJ-VERB or VERB-OBJ relation, but does not propose SUBJ-VERB-OBJ relations); second, it does not propose a clustering of occurrences, whereas distributional semantics could fill the gap and propose distributional classes at any place in the schema, instead of flat list of occurrences.

A third approach consists in tracking semantic change by applying the second aspect of the distributional hypothesis, that lexical units sharing the most of contexts are most likely to be semantically similar. This assumption has been applied to many computational semantic tasks. Applied to semantic change, if you have at your disposal a bunch of diachronic corpora, you can build the semantic vectors of any lexical unit corresponding to several periods, and track the changes from one period to another. First experiments have been proposed by (Hamilton et al., 2016). The main advantage of this approach resides in the fact that it proposes for a given word a list of semantically similar words, among which synonyms and hypernyms, which permits to clearly explicit the meaning of a word. The main drawback of this approach is to be unable to distinguish meanings for polysemous units. Another relative drawback relies on the fuzzy notion of similarity, which results in semantically too-slightly similar words (analogy), or even opposite words (antonymy). But this approach is clearly of great help to humanly grasp the meaning of a word.

In the Neoveille Project, we are currently developing a approach combining the Sketch approach mixed with semantic distributions on the main lexical unit and its arguments.

3.3.3 Tracking the Lifecycle of Neologisms

In our view, we postulate that neologisms are new form-meaning pairs (Lexical units) and thus exist from their first occurrence. Tracking the lifecycle of neologisms requires to fix criteria to identify the main phases : emergence, dissemination, conventionalization (Traugott and Trousdale, 2013). In operational systems, the main tool to follow the life of a neologism is the timeline rendering the absolute or relative frequency of the lexical unit. In Neoveille, these figures are relative to specific diastratic and diatopic paramaters, visually enabling to distinguish emergence, spread and conventionalization. These analysis are available for each identified neologism by clicking on the stats icon (see website, last neologisms menu).

4 Neoveille Tracking System Architecture

The Neoveille architecture aims at enabling a complete synergy between NLP system and expert linguists : expert linguists are not able to monitor the vast amount of textual data whereas automatic processes can help tackle this amount: experts can accurately decide if a word is or is not a neologism; our current point of view is that linguists must have the last word on what is and is not a neologism, and on the linguistic description; but as knowledge and description will grow up with time, we will build Supervised Machine Learning techniques able to predict potential neologisms.

The Neoveille web architecture has five main components:

1. A corpora manager: corpora is the main feed for NLP systems, and we propose to linguists a system enabling to choose their corpora and to add to them several meta-datas. The corpora, once defined by the user, are retrieved on a daily basis, indexed and searched for neologisms. Corpora management is available in the restricted area on the left menu;

2. An advanced search engine on the corpora; : not only corpora can be monitored, but also the system should propose a search engine with advanced capabilities : advanced querying, filtering and faceting of results; the Neoveille search engine is available on the restricted area on the left menu; based on Apache Solr, it enables to query the corpora in a multifactorial manner, with facets and visual filters;

3. Advanced Data Analytics expliciting the lifecycle of neologisms and their diachronic, di-

atopic and diastratic parameters : Neoveille provides such a Data Analytics Framework by combining meta-data to text mining; These visual analysis are available for every neologisms by clicking the stats icon;

4. A linguistic description component for neologisms : this module, whose microstructure has been setup for several years, could be used for knowledge of neology in a given language, and could also be used by a supervised machine learning system, as these features include a lot of formal properties. This component is accessible in the restricted area on the left menu.

5. formal and semantic neologisms tracking with state-of-the-art techniques The formal and semantic neologism components are accessible in the restricted area on the left menu. They work for the seven languages of the project.

5 Conclusion and perspectives

This short presentation has evoked the design and the overall architecture of a software for linguistic analysis focusing on linguistic change in a contemporary monitor corpora. It has several interesting properties :

- real-time tracking, analysis and visualization of linguistic change;
- complete synergy between Computational Linguistics processing and linguistic experts intuitions and knowledge, especially the possibility of editing automatic results by experts and exploitation of linguistic annotations by machine learning processes;
- modularity of software : corpora management, state-of-the-art search engine including analysis and visualization, neologisms mining, neologism linguistic description, lifecycle tracking.

This project, currently focusing on seven languages is in the path to extend to other languages.

References

P. Blumenthal. 2009. Éléments d'une théorie de la combinatoire des noms. *Cahiers de lexicologie 94 (2009-1), 11-29.*

Maria Teresa Cabré and Luis De Yzaguirre. 1995. Stratégie pour la détection semi-automatique des néologismes de presse. *TTR : traduction, terminologie, rédaction, 8 (2), p. 89-100.*

Emmanuel Cartier. 2016. Néoveille, système de repérage et de suivi des néologismes en sept langues. *Neologica, 10, Revue internationale de néologie, p.101-131.*

John R. Firth. 1957. Papers in linguistics 1934–1951, london, oxford university press, 1957.

Dirk Geeraerts. 2010. *Theories of Lexical Semantics*. Oxford University Press.

Paul Gevaudan and Peter Koch. 2010. Sémantique cognitive et changement sémantique. *Grandes voies et chemins de traverse de la sémantique cognitive, Mémoire de la Société de linguistique de Paris, XVIII, pp. 103-145.*

Christophe Gérard, Ingrid Falk, and Delphine Bernhard. 2014. Traitement automatisé de la néologie : pourquoi et comment intégrer l'analyse thématique? *Actes du 4e Congrès mondial de linguistique française (CMLF 2014), Berlin, p. 2627-2646.*

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *ACL 2016.*

Daphné Kerremans, Susanne Stegmayr, and Hans-Jörg Schmid. 2012. The NeoCrawler: identifying and retrieving neologisms from the internet and monitoring on-going change. *Kathryn Allan and Justyna A. Robinson, eds., Current methods in historical semantics, Berlin etc.: de Gruyter Mouton, 59-96.*

A. Kilgarriff, R. Pavel, Pavel S., and Tugwell D. 2004. The Sketch Engine. *Proceedings of Euralex, pages 105–116, Lorient.*

Antoinette Renouf. 1993. Sticking to the Text : a corpus linguist's view of language. *ASLIB Proceedings, 45 (5), p. 131-136.*

Jean-François Sablayrolles. 2016. *Les néologismes*. Collection Que sais-je? Presses Universitaires de France.

Hans-Jörg Schmid. 2015. The scope of word-formation research. *Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen and Franz Rainer, eds., Word-Formation. An International Handbook of the Languages of Europe. Vol. I.*

Elizabeth Closs Traugott and Graeme Trousdale. 2013. *Constructionalization and constructional changes.*