Towards automatic language evolution tracking A study on word sense tracking*

Nina Tahmasebi, Thomas Risse and Stefan Dietze

L3S Research Center, Hanover, Germany,
{tahmasebi, risse, dietze}@L3S.de

Abstract. Knowing about language evolution can significantly help to reveal lost information and help access documents containing language that has long since been forgotten. In this position paper we will report on our methods for finding word senses and show how these can be used to reveal important information about their evolution over time. We discuss the weaknesses of current approaches and outline future work to overcome these weaknesses.

Keywords: language evolution, word sense discrimination, clustering

1 Introduction

This work is motivated by the goal to ensure the accessibility and especially the interpretability of long term archives in order to secure knowledge for future generations. Language is evolving over time; new terms are created, existing terms change their meanings and others disappear. The available technology for accessing digital archives works well as long as the user is aware of the language evolution. But how should a young scholar find out that the term *fireman* was used in the 19th century to describe a *firefighter*?

Etymological dictionaries can be used to address the issue of language evolution by providing mappings or expanding queries. However, such dictionaries have several drawbacks. Firstly, they are rare and general. Few domain specific etymological dictionaries, such as Medline [AS05] for the medical domain, are available. Secondly, most of these dictionaries are created manually [oed,Mil95].

New kinds of digital archives and collections, e.g. Web archives, will increasingly store user generated content (e.g., Blogs, tweets, forums etc) which follow few norms. Slang and gadget names are used frequently but rarely make it into a formal dictionary. To make matters worse, these terms change at a rapid pace. Due to the change rate, as well as the huge amount of data stored in archives, it will not be possible to manually create and maintain entries and mappings for term evolution. Instead, there will be an increasing need to find and handle changes in language in an automatic way.

* This work is partly funded by the European Commission under ARCOMEM (ICT 270239)

Since automatic approaches for finding word senses within a collection of text already exist, namely word sense discrimination (WSD), these are natural starting points towards an automatic method for detecting language evolution. In [TNTR10] we presented our processing method for WSD and analyzed its applicability on historic document collections. In this paper we will focus on how WSD can be used to reveal important information about language evolution over time. We discuss the weaknesses of current approaches and outline open issues to overcome these weaknesses.

In the next section we discuss the method used for finding word senses. In section 3 we present our experiments with word sense discrimination to find language evolution. The paper finishes with conclusions and an outlook on future work.

2 Automatically Detecting Word Senses

In this paper our understanding of a word sense is to get a description of the meaning of a term in the context of the analyzed collection. In order to find word senses from large text collections, automated methods need to be exploited. For this reason we use *word sense discrimination* as an unsupervised learning method for grouping words that represent the same sense. The process consists of three main steps:

1. Pre-processing
2. Co-occurrence graph creation
3. Word sense clustering

Pre-processing We pre-process text by performing an initial cleaning of the data using regular expressions and apply an OCR error correction method described in [Nik10]. Next we extract nouns and noun phrases of size two, here on *terms*, from the cleaned text. We use two part-of-speech taggers namely TreeTagger [1] and Lingua::EN::Tagger [2] to identify and lemmatize terms. These are added to a *dictionary* corresponding to the corpora in which the terms were found.

Co-occurrence graph creation After creating the dictionary, a *co-occurrence graph* is created. All terms that are separated with an *and, or* or *comma* are considered co-occurring. For example, if the sequence "*... sports like tennis, football and rugby ...*" is found, the terms "*tennis*", "*football*" and "*rugby*" are considered co-occurring. Within the graph, each term is represented as *node* where *linked nodes* represent co-occurring terms. Finally, the graph is filtered and only co-occurrences that occur at least trice in the collection are kept. This threshold was indentified during past experiments and aims at reducing the level of noise and removing the most spurious connections.

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[2] http://search.cpan.org/ acoburn/Lingua-EN-Tagger-0.15/Tagger.pm

Word sense clustering The clustering step is the core step of word sense discrimination. The *curvature clustering algorithm* proposed by [DES04] computes the *clustering coefficient* [WS98], also called the *curvature value*, for each node to cluster the graph.

All nodes with a curvature value below a certain threshold are removed. These nodes correspond to terms that (1) have no significant sense in the collection or (2) are ambiguous, that is, they connect parts of the graph that would otherwise not be connected. By removing those terms, the graph falls apart into connected components that correspond the cluster core. E.g., the term *rock* is likely to connect terms related to its stone sense with terms from its music sense that would otherwise not be connected. To capture also the ambiguous terms, the cluster core is extended with all terms that co-occur with the terms in the cluster core. In this paper we use a curvature threshold of 0.3.

3 Towards Word Sense Evolution

3.1 Data

For our experiments we use The Times Archive [Tim08] (London) because of its long time span. The corpus consists of articles from $1785 - 1985$ and contains 7.8 million articles scanned from microfilm in 2001. The articles contain some amount of OCR errors, decreasing with time. A more in depth description of the corpus can be found in [TNTR10]. More than half of the errors were corrected during the initial cleaning of the data (Step 1 in Section 2), however, a large amount still remain. The resulting co-occurrence graphs follow the amount of errors in the data and are larger if the articles contain fewer errors. The number of clusters that can be found per year is highly dependent on the graph size and result in an average of 575 clusters per year and 7.5 terms per cluster.

3.2 Experiments

In this study we manually choose terms for which we have reason to believe there has been evolution. We look at the frequency of the terms and extract all available clusters. Our aim is to see how much can be revealed with respect to language evolution by examining word sense clusters.

St. Petersburg The city of St. Petersburg (refereed to only as Petersburg from now on) was founded in 1703 as "Sankt-Piter-Burh" and soon after renamed to "Saint Petersburg". From 1914-1924 it was named "Petrograd" and afterwards "Leningrad" and since 1991 the name is again "Saint Petersburg".

In Figure 1 the term frequencies of the city names from The Times Archive are shown. Petersburg was first mentioned in 1805 and then occasionally until 1838 after which it figured frequently in the corpus. The first mentioning of Petrograd was 1914 corresponding well to the name change. Starting 1923 the frequency of Petrograd decreases and is mentioned only occasionally after 1939. Leningrad is mentioned the first time in 1920 and then again between $1924-1985$.

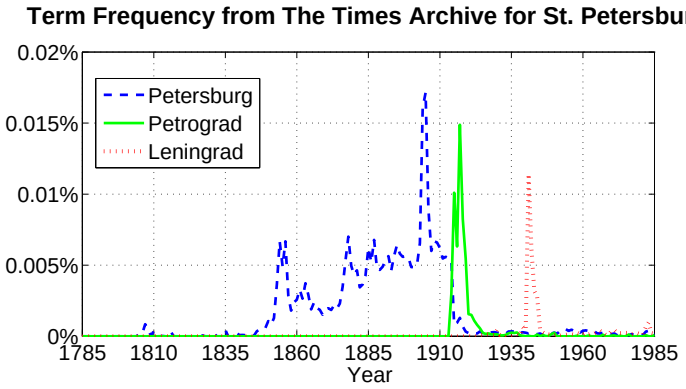

Fig. 1. Frequency of St. Petersburg, Petrograd and Leningrad in The Times Archive

year	cluster members
1856	baltic, petersburg, copenhagen, sweaborg, sevastopol, helsingfors
1886	petersburg, astrachan,moscow, caspian, sea, ararat, st, novgorod
1896	petersburg, moscow, mo, warsaw, berlin
1913	petersburg, moscow, grazing, pasture land, vladivostock
1914	russian, petrograd, nevsky, st, english bank
1915	sept, petrograd, oar, rome, budapest, in berlin,
1917	odessa, lodz, kourgan, moscow, krasnoiarsk, petrograd, koustanai
1918	petrograd, dec, july, jan, april, march, stockholm, bombay
1928	odessa, leningrad, moscow, kharkoff, nithin novgorod
1954	leningrad, moscow, kiev
1978	leningrad, moscow, kiev, novgorod

Table 1. Selected clusters and cluster members for the term *Petersburg*, *Petrograd* and *Leningrad* from The Times Archive after correction.

In Table 1 we see some clusters for Petersburg (1856-1913), Petrograd (1914-1918) and Leningrad (1928-1978). There is little in the clusters to indicate that all three terms represent the same city. However, clusters for Petrograd exist only between 1944-1918 and together with the term frequency of the term this can be seen as hints that the city of Petrograd existed only temporary. From the term frequencies we can see that Petersburg looses in frequency as Petrograd gains. Also the clusters are changed and there are no clusters for Petersburg after Petrograd has been introduced. The name change between Petrograd and Leningrad does not follow the same characteristics as the first cluster with Leningrad appears 10 years after the last one with Petrograd.

The peak in the frequencies do not offer any hints of evolution for the term. Instead the peak in 1905 for Petersburg are most likely induced by the general strike of October 1905, the peaks for Petrograd (1915-1917) and Leningrad (1941) correspond to World War I (WWI) and World War II (WWII).

Travel The term *travel* has no name change but rather a concept change. In Figure 2 we see the frequency of *travel* and *traveller* from The Times Archive. For travel we find that the frequency increases around 1912 and has a significantly higher frequency until 1985 with some dips for WWI, WWII, 1960's and 1979.

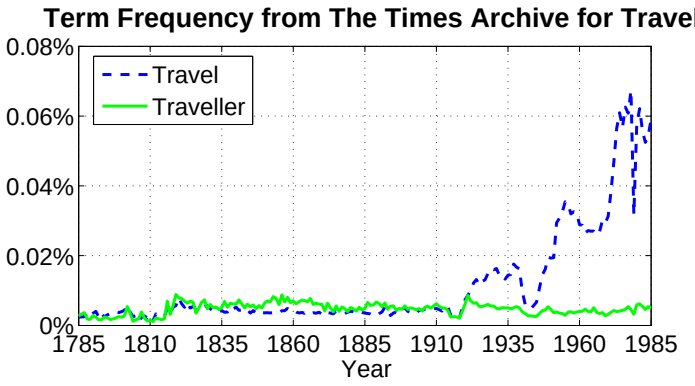

Fig. 2. Frequency of Travel and Traveller in The Times Archive

To find the sense of travel we look at Table 2 where a subset of the clusters are shown. Until 1906 we find that the term has been clustered with other terms like *literature, science, art, book* all indicating that travel was a topic reserved for the privileged few and mostly accessible in books for the rest. However, starting 1906 we find travel clustered with terms like *full board, hotel, sightseeing, sea side* to indicate that the concept of travel became more concrete and accessible in everyday life. This change coincides with a higher frequency of travel in the corpus and the clusters clearly show us that change has occurred.

year	cluster members
1803	literature, science, art, travel, voyage
1815	illustration, travels, science, travel, voyage, poetry, mile
1843	history, romance, memoir, travel, voyage, novel, biography
1867	travel, revival, colonial, foreign residence
1905	history, travel, book, mythology, biography
1906	full board, travel, best hotel
1924	town, apply, river, city, seaside, straight, travel, london, fall, country
1928	sight, meal, reserved seat, superior hotel, sightseeing, trip, travel, train
1966	loan, travel, good hotel, maintenance, fishing, tuition, hotel
1984	lanzarote, tenerife, sardinia, ravello, verona, malaga, bologna, travel

Table 2. Selected clusters and cluster members for the term *travel* from The Times Archive after correction.

A similar shift in concept can also be seen in clusters concerning *travellers*. In Table 3 we see that the type of people that traveled change. The first two clusters containing the term *yellow admiral* refer to the classic "The Wags, or the Camp of Pleasure" by Charles Dibdin. As with the senses of *travel* the traveller transforms from being a *salesman, clerk* or *merchant* to being more concrete with terms like *visa, passport, ticket, commuter*.

year	cluster members
1790	traveller, foremaft, yellow admiral, halfpay brigadier, commodore
1791	traveller, a half, yellow admiral, halfpay, brigadier, pourtrayed
1821	traveller, clothier, arrowsmith, clerk, merchant, warehouseman, silver, banker
1852	traveller, tourist, gentleman, sportsman, gamekeeper
1855	traveller, collector, clerk, cashier, bookkeeper, accountant, barmaid
1923	traveller, claiming, commercial, salesman
1932	business man, traveller, anglais, tourist, family man
1952	traveller, visa, passport, travel, ticket
1976	traveller, commuter, foreign currenc, driver

Table 3. Selected clusters and cluster members for the term *traveller* from The Times Archive after correction.

Flight The terms *aeroplane* and *aircraft* correspond to manmade devices and were introduced in The Times Archive before WWI. In Figure 3 we find the term frequencies. Both terms exhibit peaks during WWI and WWII but after WWII, aircraft gains in popularity while aeroplane is forgotten.

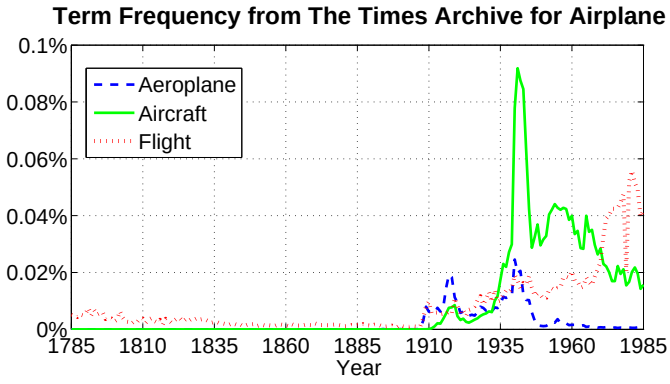

Fig. 3. Frequency of Aircraft, Aeroplane and Flight in The Times Archive

The term *flight* however, was present already before the introduction of the flying machines. In Figure 3 we see that it was present in the collection already

in 1785. Together with aircraft and aeroplane, the term flight increases in frequency before WWI. During WWII the term aeroplane is more or less replaced by aircraft. During this period, the term flight keeps a high frequency which indicates that it is related to the concept of flight and not to a specific term.

In Table 4 we can follow the evolution of the concept of flight. Between 1826-1833 the terms *robson, flight, organ builder* correspond to the names *Flight & Robson* who were indeed (church) organ builders. From 1869-1895 the clusters contain *hurdle race, flight, yard* and indicate the flight over a hurdle. 1938-1957 flight is clustered with terms like *direction, length, spin, pace* and refer to the *flight of a cricket ball*. Starting 1973 we find flight clustered with terms that represent its most common use today, a flight in a holiday sense.

year	**cluster members**
1826	robson, flight, organ, builder
1833	robson, flight, organ, builder
1869	hurdle race, flight, yard, leaving
1895	hurdle race, flight, yard, steeplechase
1938	length, flight, spin, pace, capture
1957	direction, length, spin, flight, pace
1973	flight, riding, sailing, vino, free skiing
1980	flight, visa, free board, week, pocket money, home
1984	flight, swimming pool, transfer, accommodation

Table 4. Selected clusters and cluster members for the term *flight* from The Times Archive after error correction.

3.3 Discussion

Looking at the examples presented in the previous sections, we find that they differ in character. For the St. Petersburg example, we find limited relation between the term frequencies and name changes. Instead, peaks in the frequency correspond to events. For the clusters, we also find little evidence of change. Though clusters containing a city name only exist when the city name is active, the clusters cannot directly be used to map city names automatically.

One explanation for the lack of relation can be that the clusters do not correspond to true word senses. Instead clustering algorithms aimed at capturing entity descriptions might results in clusters which can better provide a basis for finding the name changes automatically. Another possible explanation is related to the specific characteristics of individual datasets which might be more or less suitable to derive information about particular types of entities.

The travel example however, is a representative of a concept evolution rather than name change. Here we find a strong relation between increased frequency and changed meaning. Based on the flight example we recognized two aspects. Term frequency for aeroplane and aircraft appeared with the invention and introduction of the inventions in daily life. The term flight, however, changed or added

a meaning. Also, the relation between increase in frequency and the change in meaning for flight is strong. The flight example falls in the same category as *Internet* and *surfing* where Internet was the invention and surfing the term that changed/added a sense as a consequence. More in depth analysis is required to see if these relationships can be identified in an automatic fashion.

4 Conclusions and Future Work

In this study we exploited automatically identified word senses and term frequencies to investigate if language evolution could be detected. We found that concept evolution is well represented in the word senses and word sense tracking can thus be used for this type of language evolution tracking. However, word senses and frequency information were not sufficient to automatically find terms that replace each other over time (e.g., $St.Petersburg \rightarrow Petrograd$). We found that frequency bursts can be caused both by language evolution as well as events; however, event driven bursts are not presented in our clusters and need to be detected using supplementary methods. As part of future work we will focus on finding more clusters to overcome the cluster sparseness and to classify reasons for frequency bursts, e.g., strikes, fires and political events.

5 Acknowledgments

We would like to thank Times Newspapers Limited for providing the archive of The Times for our research.

References

[AS05] Andreas Abecker and Ljiljana Stojanovic. Ontology evolution: Medline case study. In *Proceedings of Wirtschaftsinformatik 2005: eEconomy, eGovernment, eSociety*, pages 1291–1308, 2005.

[DES04] Beate Dorow, Jean-pierre Eckmann, and Danilo Sergi. Using curvature and markov clustering in graphs for lexical acquisition and word sense discrimination. In *In Workshop MEANING-2005*, 2004.

[Mil95] George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38:39–41, 1995.

[Nik10] Kai Niklas. Unsupervised post-correction of ocr errors. Master's thesis, Leibniz Universität Hannover, 2010.

[oed] Oxford English Dictionary, Writing the OED. http://www.oed.com/about/writing/.

[Tim08] The Times of London, 2008. http://archive.timesonline.co.uk/tol/archive/.

[TNTR10] N. Tahmasebi, K. Niklas, T. Theuerkauf, and T. Risse. Using Word Sense Discrimination on Historic Document Collections. In *In Proc. of 10th ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Surfers Paradise, Gold Coast, Australia, 2010.

[WS98] D.J. Watts and S. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440–442, 1998.