Using Word Sense Discrimination on Historic Document Collections*

Nina Tahmasebi
L3S Research Center
Appelstr. 9a
Hannover, Germany
tahmasebi@L3S.de

Kai Niklas
L3S Research Center
Appelstr. 9a
Hannover, Germany
niklas@L3S.de

Thomas Theuerkauf
L3S Research Center
Appelstr. 9a
Hannover, Germany
theuerkauf@L3S.de

Thomas Risse
L3S Research Center
Appelstr. 9a
Hannover, Germany
risse@L3S.de

ABSTRACT

Word sense discrimination is the first, important step towards automatic detection of language evolution within large, historic document collections. By comparing the found word senses over time, we can reveal and use important information that will improve understanding and accessibility of a digital archive. Algorithms for word sense discrimination have been developed while keeping today's language in mind and have thus been evaluated on well selected, modern datasets. The quality of the word senses found in the discrimination step has a large impact on the detection of language evolution. Therefore, as a first step, we verify that word sense discrimination can successfully be applied to digitized historic documents and that the results correctly correspond to word senses. Because accessibility of digitized historic collections is influenced also by the quality of the optical character recognition (OCR), as a second step we investigate the effects of OCR errors on word sense discrimination results. All evaluations in this paper are performed on The Times Archive, a collection of newspaper articles from $1785 - 1985$.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Linguistic processing; H.3.3 [**Information Search and Retrieval**]: Clustering; H.3.7 [**Digital Libraries**]: Collection

General Terms

Algorithms, Experimentation

*This work is partly funded by the European Commission under LiWA (IST 216267)

JCDL'10, June 21–25, 2010, Gold Coast, Queensland, Australia.

Keywords

Word Sense Discrimination, Information Extraction, OCR Error Impact, Historic Document Collections

1. INTRODUCTION

The aim of word sense discrimination is to divide term collections into coherent groups of terms where each group represents one word sense or meaning. Word sense discrimination algorithms are used in many applications such as information retrieval, automatic machine translation and question answering. With historic collections, these algorithms have the potential of capturing old, as well as new meanings for a term, and hence aid in capturing language evolution. This evolution is a result of language changes that can be triggered by various factors including new insights, political and cultural trends, new legal requirements, high-impact events.

Standard information retrieval techniques cannot find relevant content created in the past since documents stored in archives might use different or outdated terms to express the sought content. A special case of evolution, outdated spellings of the same term, has been addressed in [9] where a rule based method is used for deriving spelling variations that are later used for information retrieval. In order to overcome a larger class of issues caused by language evolution in historic collections, it is necessary to develop methods and models designed especially for this purpose. Due to the size of the collections, an explicit modeling of semantics, such as those found in [9] is not possible. Therefore we use word sense discrimination as a statistical method to learn the models directly from historic archives [22]. Such models are the basis for translating the user queries of today into the terminology of the past, making the archive understandable and "accessible" to its users.

Due to the increasing efforts invested in preserving and digitizing historic documents, more and more historic collections become available in full text, e.g., The Times Archive in London, UK [3]. Beside the problems caused by language and semantics that undergo evolution, there are also issues with the digitization process, which affect the understanding of a digital archive. The digitization process needs to deal with issues such as different paper qualities, dirty

pages, different kinds of fonts or manual annotations. This causes errors in the OCR processing step which need to be handled to improve quality and readability of the archive. Unfortunately, the correction of OCR errors is often omitted due to various reasons, e.g., manual correction is expensive and time consuming while automatic correction is not fully reliable.

Existing algorithms for word sense discrimination are developed for modern language. Hence, the algorithms and the word senses provided by them are evaluated while keeping full text documents of high quality in mind. But the resulting word sense clusters that are derived using word sense discrimination on historic data, are influenced by additional factors; mainly the quality of data, e.g., OCR errors, as well as the suitability of natural language processing tools used for annotating, extracting and lemmatizing terms. The evaluation method itself can also affect the results. The behavior of these algorithms with all the issues mentioned has not yet been analyzed.

The main contribution of this paper is to evaluate the quality of word sense discrimination on historic documents when using terminology extraction and evaluation technologies for today's language. We verify that word sense discrimination can be applied on historic documents and that the resulting clusters correspond to word senses. Our analysis also measures the impact of OCR errors on the resulting word senses. As a collection we use the fully digitized archive of The Times in London, UK [3], which covers the years from $1785 - 1985$. As the digitization results undergo minor or no manual corrections, it is a good representative for evaluating the behavior of modern algorithms on historic, real world data. By verifying that the results of word sense discrimination correctly correspond to word senses, also for historic data, we take the first steps towards automatically detecting language evolution. In [21] we already presented preliminary results on a sample of the datasets. However, in this paper we extend our previous results by a deeper analysis of the whole archive.

The paper is structured as follows; the next section gives an overview of the necessary processing steps in word sense discrimination. Section 3 presents the quality measuring method and discusses possible impacts. The collection of The Times Archive that we use in our evaluation is introduced in Section 4. The evaluation results are then presented and discussed in Section 5. Afterwards in Section 6 an overview of related work is given. Finally the paper concludes and gives an outlook on future work.

2. WORD SENSE DISCRIMINATION

Word sense discrimination is the task of automatically finding the sense classes of words present in a collection. The output of word sense discrimination is sets of terms describing senses found in the collection. This grouping of terms is derived from clustering and we therefore refer to such an automatically found sense as a *cluster*. Throughout this paper we will use the terms *cluster* and *sense* interchangeably. Clustering techniques can be divided into hard and soft clustering algorithms. In hard clustering an element can only appear in one cluster, while soft clustering allows each element to appear in several. Due to the ambiguous property of words, soft clustering is most appropriate for word sense discrimination. The techniques can be further divided into two major groups, supervised and unsupervised. Because of the vast amount of data found in The Times Archive collection (s. Section 4), we are using an unsupervised technique proposed in [8], called *curvature clustering*. The curvature clustering is the core of the processing pipeline discussed in the following paragraphs.

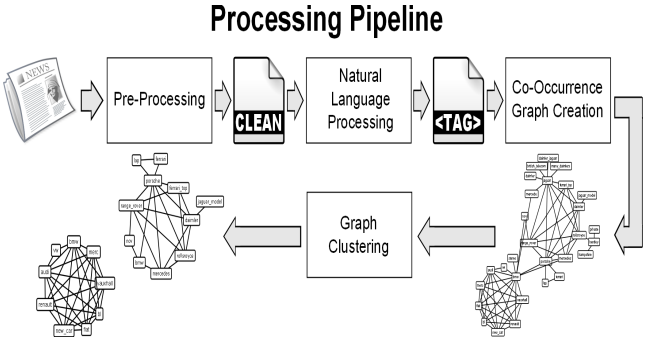

Figure 1: Overview of the word sense discrimination processing pipeline with all four steps involved.

2.1 Processing Pipeline

The processing pipeline, depicted in Figure 1, consists of four major steps; pre-processing, natural language processing, creation of co-occurrence graph and clustering. Each step is performed for each year separately. Continuing this section, we describe the processing pipeline in detail; the implementation details as well as values for thresholds are given and discussed in Section 2.2.

Pre-Processing

The first step towards finding word senses is to prepare the documents in the archive for the subsequent processing. For The Times Archive this means extracting the content from the provided XML documents and performing an initial cleaning of the data. We use a simple straightforward method for correcting OCR errors, e.g., regular expressions. More sophisticated methods can be applied to the documents at this point in the processing pipeline.

Natural Language Processing

The next step is to extract nouns and noun phrases from the cleaned text. Therefore, it is first passed to a linguistic processor that uses a part-of-speech tagger to identify nouns. In addition, terms are lemmatized if a lemma can be derived. Lemmas of identified nouns are added to a term list which is considered to be the *dictionary* corresponding to that particular year. The lemmatized text is then given as input to a second linguistic processor to extract noun phrases. The noun phrases, as well as the remaining nouns for which the first part-of-speech tagger was not able to find lemmas, are placed in the dictionary.

Co-Occurrence Graph Creation

After the natural language processing step, a *co-occurrence graph* is created. Typically the sliding window method is used for creating the graph but our initial experiments indicated that sliding windows in conjunction with the curvature clustering algorithm provide clusters corresponding to events rather than word senses. Therefore we use the following language oriented approach instead.

Using the dictionary corresponding to the particular year, the collection is searched for lists of nouns and noun phrases. Terms from the dictionary, that are found in the text separated by an "and", an "or" or a comma, are considered to be co-occurring. For example if in the sequence "*. . . cities such as Paris, New York and Berlin . . .*" the terms "*Paris*", "*New York*" and "*Berlin*" were found and assuming that they exist in the dictionary corresponding to that year, these terms are all co-occurring in the graph. Once the entire year is processed, all co-occurrences are filtered. Only co-occurrences with a frequency above a certain threshold are kept. This procedure ensures that the level of noise is reduced and most spurious connections are removed.

Graph Clustering

The clustering step is the core step of word sense discrimination and takes place once the co-occurrence graph is created. The curvature clustering algorithm by Dorow [8] is used to cluster the graph. The algorithm calculates the clustering coefficient [23], also called curvature value, of each node by counting the number of triangles that the node is involved in. The triangles, representing the interconnectedness of the node's neighbors, are normalized by the total number of possible triangles. Depicted in Figure 2 is a graph which illustrates the calculations of curvature values using different triangles. Node "*vw*" has a curvature value of 1 as it is involved in its only possible triangle "$audi, bmw, vw$", while the node "*audi*" with a curvature value of $\frac{2}{3}$ is involved in two triangles "$audi, bmw, vw$" and "$audi, bmw, fiat$" out of its three possible triangles "$audi, bmw, vw$", "$audi, bmw, fiat$", and "$audi, fiat, vw$". Node "$porsche$" is not involved in any triangle and therefore its curvature value is 0.

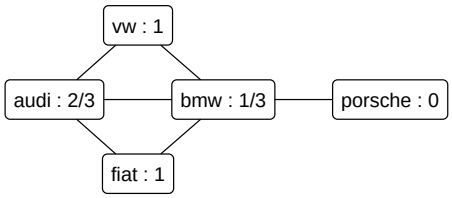

Figure 2: Graph to illustrate curvature value. Nodes are labeled with "name : curvature value".

After computing the curvature values, the algorithm removes nodes with a curvature value below a certain threshold. The low curvature nodes represent ambiguous nodes that are likely to connect parts of the graph that would otherwise not be connected. Once these nodes are removed, the remaining graph falls apart into connected components. The connected components, from now on referred to as clusters, are considered to be candidate word senses. In the final step each cluster is enriched with the nearest neighbors of its members. This way the clusters capture also the ambiguous terms and the algorithm is shown to handle both ambiguity as well as polysemy. An example will be given in Section 5.

2.2 Pipeline Implementation

When implementing the pipeline described in the previous section we rely in part on well established modules freely available. Many suitable modules are available as Perl modules and if not otherwise mentioned, Perl is used for our pipeline.

In the pre-processing step, regular expressions are used for removing non-letter characters. Dots and commas are kept because they are needed in the later steps. Dots are needed by the natural language processors for recognizing sentence structure and commas are needed for creating the co-occurrence graphs.

For the natural language processing step we use two separate processors namely TreeTagger [19] and Lingua::EN::Tagger [1]. TreeTagger is used as the first processor to find lemmas. The second processor, Lingua::EN::Tagger is used to recognize noun phrases. We restrict the length of noun phrases to length two in order to capture proper nouns like "New York".

The co-occurrence graphs are created using a Java module. Once a full co-occurrence graph corresponding to an entire year is created, it is filtered using a filtering threshold of 2, i.e., all co-occurrences with a frequency lower or equal to 2 are removed. Experiments have shown that this threshold provides good results for the majority of graphs obtained. The used threshold ensures that most of the noise is filtered out and that the resulting graphs are reasonable in size. In this paper the same threshold is applied to all graphs but the threshold can be individually learned based on the size of each graph.

For the clustering we choose the curvature threshold of 0.3 as well as 0.5. The latter has been used in previous works [7, 8, 15] and is proven to give stable word senses. Since we aim at finding word senses which evolve over time, we choose a second, lower coefficient. We thereby expect to get less strict word senses which are more likely to evolve over time. The lower coefficient should also provide us with more clusters as well as more terms in each cluster, which indicate that the clusters cover a larger portion of the collection.

3. MEASURING QUALITY

The aim of this evaluation is to measure the quality of the output provided by word sense discrimination applied on historic data. We wish to answer the question of how well the clusters found correspond to word senses. This will give insights to how well algorithms of today work without any adaptations on such datasets.

Considered Aspects

While evaluating the output of the word sense discrimination algorithm, when applied on texts older than a few decades, we need to be aware of three uncertainties which could all affect the quality of the output.

Firstly, our methods for extraction are trained on contemporary text collections thus indicating they could have difficulty recognizing terms which are no longer in use. If terms are not recognized by the natural language processors as nouns or noun phrases, they can also not participate in any co-occurrences and will therefore not be part of our clusters.

Secondly, the method for evaluation plays a role. There are several methods for evaluating clusters found by word sense discrimination algorithms [13, 16, 17, 18]. The measures can be divided into two main categories. The first uses an external source such as a dictionary or ontology for evaluation while the second relies upon a collection of sense tagged data. To our knowledge there are few or no digitized, sense tagged collections from these periods. Therefore, we must either do the tagging ourselves or use a dic-

tionary based method for evaluation. For the latter, the dictionary of choice can play a certain role. Terms that are correctly spelled, considering the time they were written, but not covered by a modern dictionary, will not be recognized as correct terms. As an example "infynyt, infinit, infinyte, infynit, infineit" are all spelling variations of the word "infinite" [2] which were correct at the time they were written, but would not be recognized by most modern dictionaries. These outdated spellings will decrease the assessed quality of the output.

Thirdly, the output is affected by the quality of the text. With a high proportion of OCR errors, terms containing errors will be recognized by neither the natural language processor, nor the dictionary used for evaluation.

Taking all the above into consideration, a low quality or quantity of clusters could indicate one of the following;

- terms are not correctly extracted by the natural language processing step because they are outdated or contain OCR errors,
- terms are not recognized by the dictionary used for evaluation, or
- the word sense discrimination algorithm used is not suitable for historic data.

When measuring the suitability of a certain word sense discrimination algorithm, all three features play a role. We intend to investigate how the results of the word sense discrimination algorithm are affected by these three uncertainties.

Method of Evaluation

For evaluating the quality of the clusters, i.e., the correspondence between clusters and word senses, we use a method proposed by Pantel and Lin [16] which relies on WordNet as a reference for word senses. The method compares the top k members of each cluster to WordNet senses. A cluster is said to correctly correspond to a WordNet sense S if the similarity between the top k members of the cluster and the sense S is above a given threshold. Following [16] we choose similarity threshold 0.25 and set the number of top k members to $k = 4$. The clustering algorithm proposed in [16] assigns to each cluster member, a probability of belonging to that cluster, thus providing an intuitive way of choosing "top" members. The curvature clustering algorithm does not provide such probabilities and therefore we choose our k members randomly.

4. THE TIMES ARCHIVE

For evaluation we use The Times Archive [3] and in this section we provide an overview of the dataset. The Times Archive consists of news paper articles spanning from year 1785 to 1985. The digitization process was started in year 2000 when the collection was digitized from microfilm and OCR technology was applied to process the images. The resulting 201 years of data, each year considered as a separate dataset, consists of 4363 articles in the smallest dataset and 91583 in the largest. The number of whitespace separated tokens range from 4 million tokens in 1785 to 68 million tokens in 1928. In sum we found 7.1 billion tokens that translate into an average number of 35 million tokens per year.

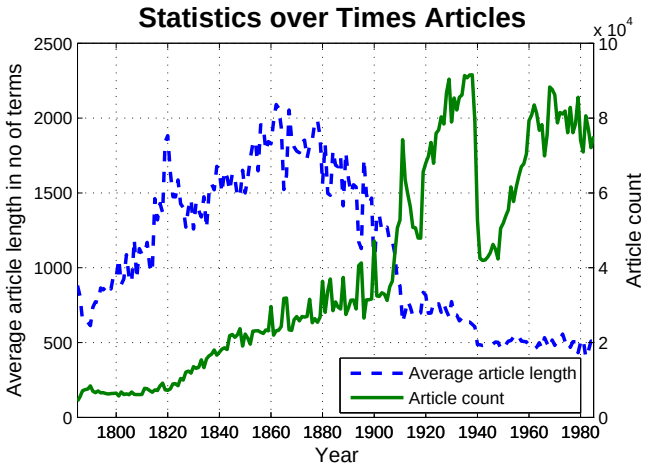

Figure 3: Number of articles and average length of articles in The Times Archive from 1785 **to** 1985**.**

Starting with almost 4400 articles from 1785, the number of articles increase steadily during the first 100 years as shown in Figure 3. In the early 20th century the increase becomes more rapid and in 1911 we have almost double the number of articles as in 1905. The higher number of articles is affected during World War I (WWI) and World War II (WWII). In fact, in both periods the number of articles decreases heavily. The maximum number of articles is found in year 1938 when almost 92000 articles are published.

When considering the length of an article, we count the number of terms in the article. A term is a space separated single word. We find that the average length of articles increase from 1785 until 1862 when a maximum of almost 2100 terms per article is measured. After follows a period of decrease which continues until 1940, then the average length of articles converges at roughly 500 terms per article.

4.1 Article Categories in The Times

All articles in The Times Archive were manually classified into 18 categories during the digitization process. Overall it can be observed that some categories are stable over time while others gain or lose in popularity. We measure popularity for each category in average number of articles per year. Among the categories, the "*News*" category is not surprisingly the most stable category over time. Categories "*Sport*" and "*Obituaries*" are examples that gain in popularity while "*Politics and parliament* " as well as "*Birth, Marriages and Deaths*" are examples that lose in popularity over the years.

In Figure 4 we can clearly see that the "*News*" category dominates in popularity. Roughly $30\% - 40\%$ of all articles are classified as news each year. The category "*Sport*" shows an increase over the years which clearly dips during both world wars. The "*Birth, Marriages and Deaths*" category loses in popularity over the years but unlike the category "*Sport*", it peaks, not very surprisingly, during both world wars.

4.2 OCR Quality

To better understand the output of the word sense discrimination algorithm, we need to measure the distribution of OCR errors in the collection over time. An analysis of

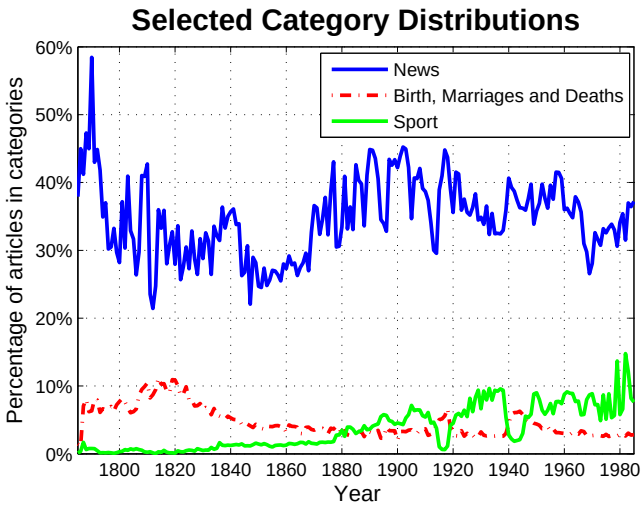

Figure 4: Distribution of articles classified under categories "*News*", "*Birth, Marriages and Deaths*" and "*Sport*" in The Times Archive.

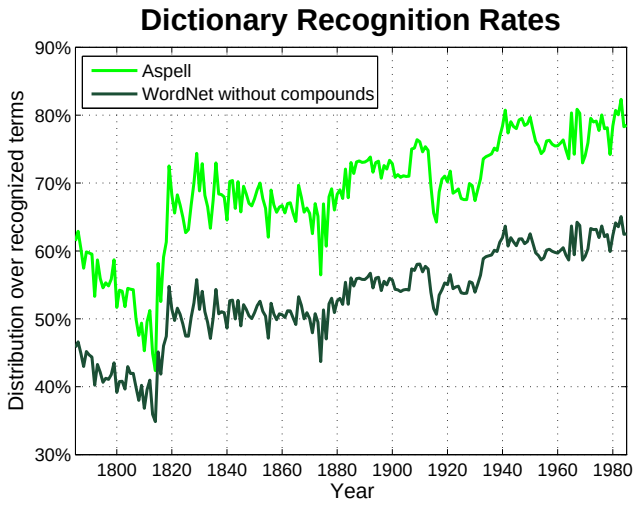

Figure 5: Percentages of terms covered by the dictionaries WordNet and Aspell each year.

The Times Archive shows three main types of OCR errors, often a mixture of these is the case.

- *Segmentation errors.* Different line, word or character spacings lead to misrecognitions of white spaces causing segmentation errors (e.g. "thisis", "depa rtmen t").
- *Syllable divisions.* Words are split up with line breaks if they are too long, which increase the number of segmentation errors (e.g., "de- partment").
- *Misrecognition of characters.* Dirt and font-variations prevent an accurate recognition of characters which induce wrong recognitions of words (e.g., "souiid", "&-Bi1rd#!").

The amount of OCR errors is approximated using a *dictionary recognition rate*. The dictionary recognition rate measures the portion of the language which is covered by a modern dictionary. The OCR errors are considered to be $OCRerror \approx 1 - f(t)$ where $f(t)$ is the dictionary recognition rate for a given dictionary (such as Aspell and WordNet) and a time period t (in our case a year). Outdated terms can lack OCR errors but still not be recognized by the dictionary. We therefore consider this approximation, in addition to OCR errors, to also capture outdated terms. The text is cleaned and run through two separate dictionaries, one token at the time. Cleaning refers to removing heading and trailing non-letter characters while leaving any characters in a term, e.g., "&Bi1rd#!" becomes "Bi1rd".

The dictionaries used are extracted from WordNet 3.0 [14] and GNU Aspell 0.60.6 [5], here on referred to as WordNet and Aspell. WordNet contains about $147k$ unique single as well as compound terms but no stopwords. Aspell contains roughly $138k$ unique terms without any compound terms. Since we run each term separately through the dictionaries, we disregard compound terms in WordNet which leads to a reduced size of roughly $83k$.

Due to the fact that WordNet only contains lemmas we have to lemmatize each token before we can check it. Therefore we use the stemmer from the MIT Java WordNet interface JWI [11] which follows WordNet's stemmer implementation with one additional rule for terms ending on "-ful". Additionally, we add WordNet's "exception entries" to the dictionary. These entries contain mappings from irregular words to its corresponding lemmas which the stemmer cannot compute. Including these entries the WordNet dictionary contains about 89k terms. The Aspell dictionary contains not only lemmas but additionally morphologies and names. Therefore no lemmatization and "exception entries" are necessary.

It can be seen from Figure 5, that the two dictionaries differ in coverage. While Aspell covers from $42\% - 82\%$ of all terms in our collection, WordNet ranges from $35\% - 65\%$. On average, Aspell covers 69% whereas WordNet only covers 53% of all terms. When adding stopwords, WordNet displays almost the same mean and variation as Aspell.

The recognition rates found by Aspell decrease from 61% near too nearly 42% between 1785 to 1814. After 1814 the recognition rate increases steeply to between $63\% - 74\%$ before it finally settles around $75\% - 80\%$ around mid 20th century. The large difference between 1814 and 1815 found for both dictionaries, is caused by the introduction of a steam press in end of 1814 [4]. The decrease in quality from 1785 until 1815 is likely caused by the logographic printing blocks used for printing during the period. They wore out quickly and had to be replaced often. There is an editorial in The Times addressing this issue with an apology and a promise to attend to the problem.

In a complementary analysis, which is not shown as a figure in this paper, we measure the portion of terms recognized as nouns by WordNet as we need them later for the graph creation. We find that this ranges from slightly above 31% to 53%. To measure the suitability of TreeTagger as a lemmatizer we measure the proportion of WordNet nouns for which TreeTagger found a lemma. This is a steadily increasing number which ranges from $57\% - 67\%$. This means that at best, TreeTagger cannot find lemmas for one third of all terms recognized as nouns by WordNet.

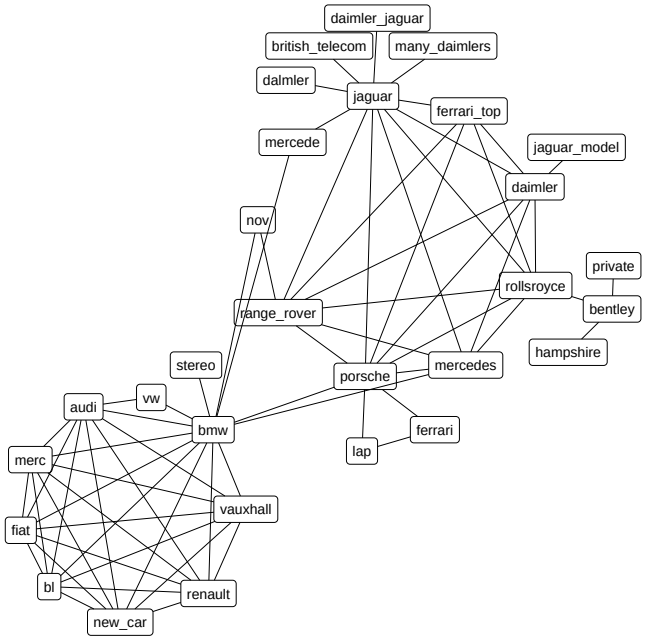

Figure 6: Subgraph from the year 1985 including two clusters representing sports cars and family cars.

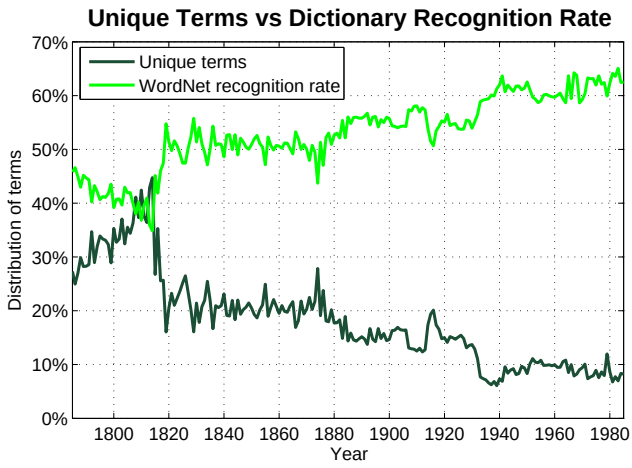

Figure 7: Percentage of unique terms in the collection compared to the dictionary recognition rate based on WordNet.

5. EVALUATION RESULTS

In the following section we present and discuss the results we have gathered in our evaluation. We first analyze the potential impact of OCR errors on the word sense discrimination results. In Section 5.2 we continue with statistics regarding the clusters which are necessary for the final quality evaluation in the section following. Finally in Section 5.4 we discuss the results and derive our conclusions.

In Figure 6 we see a sample output of a co-occurrence graph extracted from the graph created using the 1985 collection. The corresponding clusters are listed below. Due to space issues not all members of the clusters are shown in the graph.

1. bentley, jaguar, range rover, porsche, bmw, jaguar model, lap, daimler ,bmws, rolls royce, ferrari top, company, nov, ferrari, mercedes

2. renault, vauxhall, merc, w, nb, audi, golf, e, honda, volvo, fiat, b t, nb, vw, ford, ib, te, wl, audi, bmw, new car, opel, bl

Both clusters correspond to cars, though they are different in that the first cluster represents fast, expensive sports cars while the other represents every day, family cars. While the first cluster mostly contains terms which make sense together, the second cluster shows a higher level of noise. This noise is a result of the pre-processing step as well as the natural language processor step. "nb, b t, ib" etc. are incorrectly identified nouns or noun phrases by the natural language processor, though many of these should be considered as OCR errors.

5.1 Impact of OCR Errors

In digitized collections OCR errors are an obvious reason for having a large number of unique terms. Therefore, we analyze the relation between OCR errors and the number of unique terms and investigate the implications this has on the output of the word sense discrimination algorithm.

In Figure 7 we compare the percentage of unique terms against the WordNet recognition rate, the analysis is analog for Aspell recognition rate. Considering the formula for $OCRerror$ from Section 4.2 we note that the graphs look like inverses of each other. In the first period, $1785-1814$, WordNet covers a decreasing amount of terms while during the same period the percentage of unique tokens increases. The period $1820-1880$ corresponds to a rather stable rate of unique terms as well as terms recognized by WordNet. The peak in number of unique terms year 1874 corresponds to the dip in the dictionary recognition rate for the same year. This indicates a high amount of OCR error since a 10% increase of new terms in the newspaper within one year, which disappear again in the next, seems extremely unlikely. The peak which occurs during $1914-1918$ corresponds to WWI and again the increase in unique terms correspond to a dip in dictionary recognition rates for the same period. We can conclude that also this period is affected by many OCR errors.

After WWII both graphs are relatively stable and the percentage of unique terms deviates between $7\%-12\%$. For the peak during WWI in number of unique tokens, one possible explanation could be a higher rate of names of fallen soldiers which affect the amount of unique terms. The fact that we do not experience the same behavior in the period for WWII renders the explanation unlikely.

After concluding that year 1874 and the period of WWI are likely to have a high percentage of OCR errors, we investigate how this affects the clusters. We find that the number of clusters dramatically decreases during these periods in comparison to the neighboring years, e.g., in year 1873 and 1875 there are 348 and 579 clusters respectively, in 1874 there are merely 91 clusters.

5.2 Cluster Analysis

Applying the method described in Section 2 to The Times Archive results in $221-106000$ unique relations, i.e., edges in the graph. The number of unique relations per year is highly correlated to the number of nouns recognized by WordNet

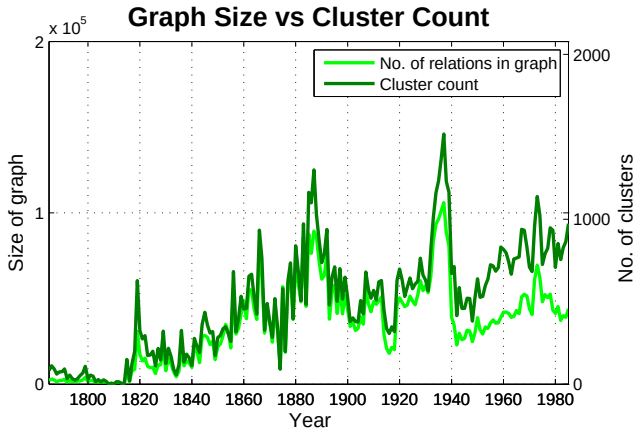

Figure 8: Relation between the size of the graph and the number of clusters found with curvature value 0.3 for each year.

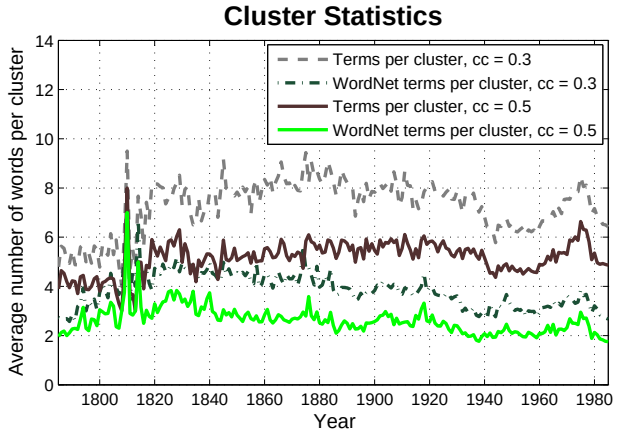

Figure 9: Average cluster sizes for each year w.r.t. term count, WordNet term count and different curvature values.

for that year. Figure 8 shows the relation between graph sizes and corresponding number of clusters. It can be seen that the number of clusters depend on the number of relations. For curvature value 0.5 the number of clusters are in average 38% less than for a curvature value of 0.3 but both follow the same distribution.

After WWII we find that the number of clusters found w.r.t the size of the graph, increases. This indicates that the curvature clustering algorithm performs better w.r.t the quantity of found clusters in this period. It is interesting to note that this coincides with the period of low percentage of unique terms (s. Figure 7). As with the number of clusters, the total number of words in all clusters differs between the two curvature values. As expected, the lower value of 0.3 covers more terms over the entire period, i.e., the clustering algorithm produces more terms in each cluster in comparison to the higher curvature value.

The average number of terms in each cluster is depicted in Figure 9. It can be observed that both curvature values behave similar to each other but differ in value. We also measure the average number of terms in each cluster, which can be found in WordNet, here on called *WordNet terms*. We note that the average number of terms in total and the average number of WordNet terms behave differently. The latter number decreases however slowly, starting early 19th century. This indicates that each cluster contains fewer WordNet terms, even though the portion of WordNet nouns in the collection increases over the year as reported in Section 4. One explanation for this is that the percentage of clusters which contain no WordNet terms increase from $0\% - 20\%$ over the entire period. This could indicate that the language becomes less strict in The Times over the years.

For the average number of terms as well as WordNet terms a spike can be seen in years $1808 - 1810$. The reasons for this spike are very low numbers of clusters found in the period, e.g., in year 1810 for curvature value 0.3 there are two clusters and for value 0.5 there is only one cluster. Some examples of clusters are available in Appendix.

5.3 Cluster Quality Evaluation

Starting 1785, a sample of 4 years of data every 50 years is chosen for evaluation of the cluster quality (s. Section 3). This 10% sample is used to get an overview of the word sense discrimination algorithm. The cluster evaluation is performed on the clusters created for each year, using the curvature value of 0.3. The results are shown in Figure 10. We measure precision for each year as the proportion of clusters that correctly correspond to a WordNet sense.

The minimum precision for a year is 68% measured for 1886. The maximum precision is 91% corresponding to the cluster set of 1785. A student t-test with $\alpha = 0.1$ shows that the mean precision for the first two periods is higher than the mean precision for the last three periods. This is likely highly connected with the lower amount of clusters with WordNet terms for the last three periods.

Comparing these numbers with the ones originally presented in [21] for the same 10% sample, we find that the precisions obtained here are slightly lower. The analysis in [21] took only into account nouns but no noun phrases, leading to much smaller graphs and a decreased number of clusters. When considering noun phrases, there are many people names and places, e.g., “Mr. Alfred”, “south Yemen”, which are not covered by WordNet and hence decrease the quality. In [21] a maximum of 180 clusters for any one year was reported while we now have a maximum of nine times as many. We omitted an evaluation of the clusters created with curvature value 0.5 since they were shown to have an equal or lower precision than the corresponding clusters for curvature value 0.3.

5.4 Discussion

Overall the results show that the used word sense discrimination algorithm can be applied to historic documents dating back at least to 19th century. However, the results are influenced by various factors which can be seen from the figures.

Before starting these evaluations we worked under the hypothesis that more terms would be covered the more recent the texts. As a result of this we assumed that the quality of clusters produced would increase over time, starting at a very low quality. While the results supported the first hypothesis, we could find evidence which contradicted our assumption about the quality of the clusters. Instead we find that the clusters found during the earlier period keep

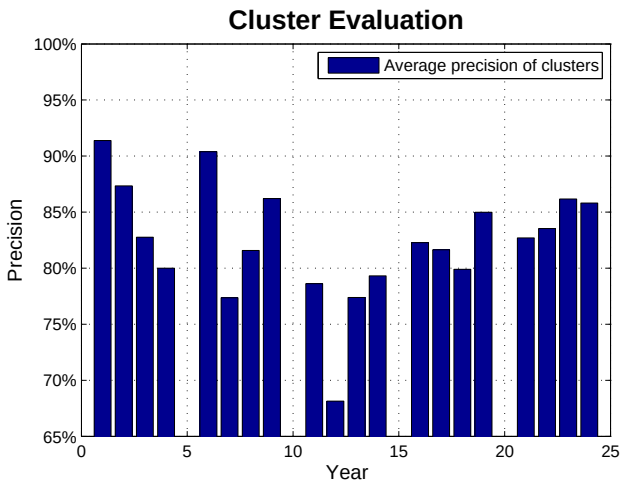

Figure 10: The results of the cluster evaluations on a 10% sample for clusters with curvature value 0.3.

a high quality. This could be explained by the fact that a very high percentage of these clusters contain WordNet terms. The coverage on the other hand for the same period is much lower; we have fewer clusters and fewer terms in each cluster. This indicates that while our methods cover a much smaller portion of the data during the first 50 years, the covered portion keeps a high quality.

The analysis of unique terms versus the dictionary recognition rate gives us some insight to the effects of OCR errors for word sense discrimination (s. Figure 7). We observe that there is a higher rate of unique terms during the period of $1815-1880$ compared to the period $1950-1985$. This corresponds to fewer words recognized by the WordNet dictionary as well as Aspell during $1815-1880$ compared to the latter period. This leads to one of the following;

1. *There were more terms in use during the 19th century.* It could also be that The Times had a writing style with a more varied language at that time. If so, the lower recognition rate implies that the higher amount of unique terms correspond to outdated terms.

2. *There were more OCR errors during the first period.* It could be that better paper quality and printing technology over time as well as lower storage time result in a decrease of the OCR error rate over time.

3. *A mixture of the above.* The higher percentage of terms that could not be recognized during the first period can be due to both of the above stated; terms which contain OCR errors as well as terms which are outdated.

Though it is likely that 2. is the most dominating reason for the increased number of unique term, further studies must be conducted to give a fully satisfying answer.

We note that the variations in the recognition rates are very similar between the two dictionaries. This leads us to the conclusion that while the coverage of the dictionary does play a certain role, the variations in these rates depend highly on OCR errors present in the texts. We conclude that the quality of the clusters produced is not significantly affected by the variation in the dictionary recognition rates while the coverage is highly affected. For the periods where we have larger dips (1814, 1874 and $1914-1918$) we have reason to assume that there are high rates of OCR errors as we have a significant decrease in the number of clusters compared with neighboring years.

Also affecting the output of the word sense discrimination algorithm is the number of nouns recognized as well as the number of those that can be lemmatized. We observe a high correlation between the number of lemmatized nouns found in a year and the size of the graph corresponding to that year. Therefore it is very important to find better natural language processing tools covering also historic texts. The current level of lemmas found among the nouns, a maximum of 67%, is not sufficient for this purpose.

The method that we have chosen for word sense discrimination has shown to give higher quality clusters while having a lower coverage. To increase the coverage we use nouns as well as noun phrases with a length of 2 during the co-occurrence graph creation. Currently we investigate if additional patterns will help to capture more relations, such as hypernym- and meronym ("IS-A" and "PART-OF") relations. Once these relations have been found, they can be added to the co-occurrence graph before we do the clustering, or used to extend the clusters after the clustering step. We intend to investigate if these additional patterns, found in the works of Hearst [12], could lead to more and larger clusters for older texts. We also want to investigate if the added patterns can improve the quality of the clusters. If these patterns are not appropriate for older texts, the next step is to learn patterns automatically. Another direction is to add verb contexts in the co-occurrence graph [16]. The clusters would contain only nouns, while taking into account the additional information provided by verb contexts.

6. RELATED WORK

In the field of word sense discrimination, as well as word sense disambiguation, it is common to evaluate the algorithms on digital collections covering documents created in the 2nd half of the 20th century. These collections are error free and sometimes annotated with linguistic annotations. Except our preliminary work in [21], an evaluation of word sense discrimination on document collections covering more than 50 years, has to our knowledge not been performed so far. The impacts of OCR errors on word sense discrimination has also not been previously investigated. Therefore we focus our discussion on word sense discrimination algorithms and evaluation methods.

Several methods for word sense discrimination based on co-occurrence analysis and clustering have been proposed like [6, 17, 20]. Schütze [20] presented the idea of context group discrimination. Each occurrence of an ambiguous word in a training set is mapped to a point in word space. The similarity between two points is measured by cosine similarity. A context vector is then considered as the centroid (or sum) or the vectors of the words occurring in the context. This set of context vectors are then clustered into a number of coherent clusters. The representation of a sense is the centroid of its cluster.

The use of dependency triples is one alternative approach for word sense discrimination and was first described in [13]. In this paper a word similarity measure is proposed and an automatically created thesaurus which uses this similarity

is evaluated. This method has the restriction of using hard clustering which is less appropriate for word senses due to ambiguity and polysemy of words. The author reports the method to work well but no formal evaluation is performed. In [16] a clustering algorithm called Clustering By Committee (CBC) is presented, which outperforms popular algorithms like Buckshot, K-means and Average Link in both recall and precision. The paper proposes a method for evaluating the output of a word sense clustering algorithm to WordNet, which has since been widely used [7, 10]. In addition, it has been implemented in the WordNet::Similarity package by Ted Pedersen et al. [17]. Due to a wide acceptance of the method, we based our methods of evaluation on this work.

Dorow et al. [7, 8] presented an another method for taking semantic structures in to account in order to improve discrimination quality. They showed that co-occurrences of nouns in lists contain valuable information about the meaning of words. A graph is constructed in which the nodes are nouns and noun phrases. There exists an edge between two nodes if the corresponding nouns are found separated by "and", "or" or commas in the collection. The graph is clustered based on the clustering coefficient of a node and the resulting clusters contain semantically related terms representing word senses. The method can handle ambiguity and due to the good results reported in [7, 8] we have decided to use this method for our processing pipeline.

7. CONCLUSIONS & FUTURE WORK

We have investigated whether current word sense discrimination algorithms can be applied on historic document collections. Because many digitized collections contain OCR errors, we investigated the effects of OCR errors on the word senses found. For our evaluations we used 201 years of newspaper articles from The Times Archive. We conclude that the chosen algorithm works well over the entire collection. The clusters produced for 18th and 19th century correspond well to word senses. Though the clusters are of high quality, we found that the number of clusters is highly related to the amount of OCR errors. Furthermore we found that natural language processing tools for recognizing part-of-speech and lemmatizing terms must be improved for high quality processing of historic data.

However, based on the results presented, we conclude that the found word senses can be used as a basis for finding language evolution by tracking the evolution of word senses.

As a next step towards detecting language evolution we intend to improve the quality and quantity of word senses. This will be done by automatically correcting OCR errors, using additional patterns as well as investigating the possibility of adding verb contexts for creating co-occurrence graphs.

8. ACKNOWLEDGMENTS

We would like to thank Times Newspapers Limited for providing the archive of The Times for our research. We would also like to thank Beate Dorow for sharing code and insights.

9. REFERENCES

[1] Lingua::EN::Tagger - part-of-speech tagger for english natural language processing. `http://search.cpan.org/~acoburn/Lingua-EN-Tagger-0.15/Tagger.pm`.

[2] Oxford English Dictionary. `http://www.oed.com/`.

[3] The Times Archive. `http://archive.timesonline.co.uk/tol/archive/`.

[4] The Times, November 29, 1814. `http://archive.timesonline.co.uk/tol/viewArticle.arc?articleId=ARCHIVE-The_Times-1814-11-29-03-003&pageId=ARCHIVE-The_Times-1814-11-29-03`.

[5] K. Atkinson. GNU Aspell version 0.60.6, Released under the GNU LGPL license in April, 2008. `http://aspell.net/`.

[6] D. Davidov and A. Rappoport. Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 297–304, Sydney, Australia, 2006.

[7] B. Dorow. *A Graph Model for Words and their Meanings*. PhD thesis, University of Stuttgart, 2007.

[8] B. Dorow and D. Widdows. Discovering corpus-specific word senses. In *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, pages 79–82, Budapest, Hungary, 2003.

[9] A. Ernst-Gerlach and N. Fuhr. Retrieval in text collections with historic spelling using linguistic and spelling variants. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 333–341, Vancouver, BC, Canada, 2007. ACM.

[10] O. Ferret. Discovering word senses from a network of lexical cooccurrences. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, 1326, Geneva, Switzerland, 2004.

[11] M. A. Finlayson. MIT Java Wordnet Interface version 2.1.5, Released under Creative Commons Attribution-NonCommerical Version 3.0 Unported License. `http://projects.csail.mit.edu/jwi/`.

[12] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545, 1992.

[13] D. Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational Linguistics*, pages 768–774, Montreal, Quebec, Canada, 1998.

[14] G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995.

[15] G. Palla, A.-L. Barabasi, and T. Vicsek. Quantifying social group evolution. *Nature*, 446:664–667, 2007.

[16] P. Pantel and D. Lin. Discovering word senses from text. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619, Edmonton, Alberta, Canada, 2002. ACM.

[17] T. Pedersen and R. Bruce. Distinguishing word senses in untagged text. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 197–207, Providence, RI, 1997.

[18] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal, Quebec, Canada, 1995.

[19] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK, 1994. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`.

[20] H. Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, 1998.

[21] N. Tahmasebi. Automatic detection of terminology evolution. In *OTM Workshops*, pages 769–778, 2009.

[22] N. Tahmasebi, S. Ramesh, and T. Risse. First results on detecting term evolutions. In *9th International Web Archiving Workshop*, Corfu, Greece, 2009.

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

APPENDIX

A. CLUSTER EXAMPLES

Due to repetitions, the clusters shown in the Appendix are sampled from all clusters mentioning each term and a limited number of terms are shown for each cluster. In both cluster sets we find that the number of terms in each cluster increases over time. It should be clear that clusters displayed here do not follow the evolution of each term as a whole, but as it was mentioned in The Times Archive.

In Table 1 we see clusters for the term *flight*. Among the displayed clusters it is clear that the senses for flight are several and mostly grouped together. Between 1867-1894 there are 5 clusters (only two of them displayed here) that all refer to hurdle races. Between the years 1938 - 1957 the clusters are referring to cricket, the terms in the clusters are referring to the ball. Starting from 1973 the clusters correspond to the modern sense of flight as a means of travel, especially for holidays. The introduction of among others *pocket money, visa, accommodation*, differentiates the latter clusters from the earlier. Also the cluster in 1927 refers to a flight but not necessarily in a holiday sense.

year	cluster members
1867	yard, terrace, flight
1892	hurdle race, flight, year, steeplechase
1927	flight, england, london, ontariolondon
1938	length, flight, spin, pace
1957	flight, speed, direction spin, pace
1973	flight, riding, sailing, vino, free skiing
1980	flight, visa, free board, week, pocket money, home
1984	flight, swimming pool, transfer, accommodation

Table 1: Selected clusters and cluster members for the term 'flight'.

In Table 2 we show a set of clusters corresponding to the term *mechanic*. We begin by noting that because the terms are lemmatized, we cannot distinguish between mechanic as a craftsman and mechanics as a science. This can be seen in the table in that the clusters are a mix of both senses. While the cluster in 1852 represents the occupation, the cluster in 1891 is clearly describing mechanics as a science. The introduction of *electricity* with mechanic came first in 1891. It is also interesting to note that the term *software* appeared together with mechanic in 1985.

year	cluster members
1818	chemistry, philosophy, mechanic
1829	optical instrument, optic, mechanic
1852	labourer, tradesman, artisan, mechanic
1880	german, animal physiology, mechanic, artisan
1891	magnetism, physics, mechanic, science, electricity, physiology, astronomy, mathematics, ...
1963	electrical worker, enginemen, mechanic, fireman
1974	atomic structure, play school, art, applied calculus, mechanic, quantum theory
1985	tooling, software, electronics, mechanic

Table 2: Selected clusters and cluster members for the term 'mechanic'.