Search Result Diversification Methods to Assist Lexicographers

Lars Borin Markus Forsberg Karin Friberg Heppin
Richard Johansson Annika Kjellandsson
Språkbanken, Department of Swedish, University of Gothenburg
Box 100, SE-40530 Gothenburg, Sweden
first.last@svenska.gu.se

Abstract

We show how the lexicographic task of finding informative and diverse example sentences can be cast as a search result diversification problem, where an objective based on relevance and diversity is maximized. This problem has been studied intensively in the information retrieval community during recent years, and efficient algorithms have been devised. We finally show how the approach has been implemented in a lexicographic project, and describe the relevance and diversity functions used in that context.

1 Introduction

Modern lexicography is empirical: the lexicographer describing a word, phrase, or construction needs to understand its variations in patterns of usage by searching in large and diverse set of corpora to see the contexts in which it appears (Atkins and Rundell, 2008). Unless studying very rare phenomena, it is then important that the lexicographer has access to usable tools that are able to search in a corpus and quickly aggregate the results in a way that is meaningful for the lexicographic task at hand. The results of this aggregation can then be used when selecting example sentences for inclusion in dictionary entries.

What kind of aggregation would a lexicographer need? As we have hinted above, the goals are twofold: 1) selection of representative and relevant prototypes; 2) giving a good overview of the diversity of the full search result. There are a number of automatic methods for selection of examples for lexicographers, most of which have focused on the first of these goals. One well-known method is GDEX (Kilgarriff et al., 2008), which has been used in conjunction with the Sketch Engine (Kilgarriff et al., 2004) in several lexicographic tasks. GDEX uses a set of rules of thumb designed to address the relevance issue for lexicographers: example sentences should be medium-short (but not too short) and avoid rare words and syntactic constructions, and the search term should preferably be in the main clause.

In this paper, we argue that the two goals of representativeness and diversity can be cast as a *search result diversification* problem. The task of diversification has seen much recent interest in the information retrieval community (Gollapudi and Sharma, 2009; Drosou and Pitoura, 2010). While diversification is computationally intractable in most cases, fast approximation algorithms exist (Drosou and Pitoura, 2009; Minack et al., 2011) and have facilitated the development of practical systems for the diversification of search results for searches on the web, for documents as well as images (Hare et al., 2009; Krestel and Dokoohaki, 2011). Note that the purpose of diversification in information retrieval is typically different from that in lexicography: increasing the probability of finding a particular piece of information that the user is looking for.

2 Diversification of Search Result Sets

We will now formally define the problem of set diversification (Drosou and Pitoura, 2010). We assume that we are given a *relevance function* $r(i)$ that assigns a "suitability" score to an item i, and a *distance function* $d(i, j)$ that measures how different the two items i and j are. These functions should be tailored to suit the task at hand.

Assuming we are looking for a subset of size k of a full set $\boldsymbol{U}$ of search results. Then for given relevance and distance functions r and d, we define

Proceedings of the 6th Linguistic Annotation Workshop, pages 113–117,
Jeju, Republic of Korea, 12-13 July 2012. ©2012 Association for Computational Linguistics

the diversification task as an optimization problem where we find the subset $\boldsymbol{S}_k^*$ that maximizes some objective f:

$$\boldsymbol{S}_k^* = \underset{\substack{\boldsymbol{S}_k \subseteq \boldsymbol{U} \\ |\boldsymbol{S}_k| = k}}{\arg\max} f(\boldsymbol{S}_k, r, d)$$

How should we then choose the objective f in terms of the relevance r and distance d? One obvious way is to sum all relevance and pairwise internal distance scores. This objective is called the SUM function.

$$f_{\text{SUM}}(\boldsymbol{S}_k, r, d) = (k-1) \sum_{i \in \boldsymbol{S}_k} r(i) + \lambda \sum_{\substack{i,j \in \boldsymbol{S}_k \\ i \neq j}} d(i,j)$$

Here λ is a weight controlling the tradeoff between relevance and distance.

Another possible objective, the MIN function, uses the minimum relevance and internal distance:

$$f_{\text{MIN}}(\boldsymbol{S}_k, r, d) = \min_{i \in \boldsymbol{S}_k} r(i) + \lambda \min_{\substack{i,j \in \boldsymbol{S}_k \\ i \neq j}} d(i,j)$$

The problems of finding the sets maximizing these objectives are referred to as MAXSUM and MAXMIN, and they are both NP-hard and need approximations to be usable in practice.

2.1 Approximate Diversification of Search Result Streams

There are a number algorithms to solve the MAXSUM and MAXMIN optimization problems approximately (Drosou and Pitoura, 2009). In this paper, we will make use of the online diversification algorithm presented by Minack et al. (2011). This algorithm is completely incremental, which leads to several advantages: 1) the processing time is linear in the number of search hits, as opposed to other algorithms that have higher computational complexity; 2) we do not have to know the size of the full result set beforehand; 3) we do not have to keep the full set in memory; 4) intermediate results are meaningful and can be presented to the user, which improves the feeling of responsiveness of the user interface. Minack et al. (2011) found that the greedy approximation algorithm produced diverse subsets of a quality comparable to that of more complex algorithms. However, one question they did not address is how the efficacy of the greedy algorithm is affected by the properties of the relevance and distance functions.

The incremental diversification algorithm is very simple. A diverse set $\boldsymbol{S}$ is maintained at each step, and when we encounter a new item i, find the item j in the current instance of $\boldsymbol{S}$ that leads to the maximal increase in f when adding i and removing j. This means that we enforce the size constraint of $\boldsymbol{S}$ at all times. Algorithm 1 shows the pseudocode.

Algorithm 1 Diversification of a stream of search results (Minack et al., 2011).

```
input Search result iterator $I$
      Maximum size $k$ of the output set
      Optimization objective function $f$
$\boldsymbol{S} \leftarrow \emptyset$
while $I$ has another item $i$
  if $|\boldsymbol{S}| < k$
    $\boldsymbol{S} \leftarrow \boldsymbol{S} \cup i$
  else
    $\boldsymbol{S}_{max} \leftarrow \boldsymbol{S}$
    for $j$ in $\boldsymbol{S}$
      $\boldsymbol{S}' \leftarrow \boldsymbol{S} \cup \{i\} \setminus \{j\}$
      if $f(\boldsymbol{S}', r, d) > f(\boldsymbol{S}_{max}, r, d)$
        $\boldsymbol{S}_{max} \leftarrow \boldsymbol{S}'$
    $\boldsymbol{S} \leftarrow \boldsymbol{S}_{max}$
 return $\boldsymbol{S}$
```

We omit the description of further implementation details. In particular, the f_{SUM} and f_{MIN} objectives can be computed by incremental updates, which speeds up their evaluation greatly.

3 A Case Study: Diversity and Relevance in a Lexicographic Project

We applied the search result diversification method in a new annotation user interface used in the Swedish FrameNet (SweFN) project. This is a lexical resource under development (Borin et al., 2010; Friberg Heppin and Toporowska Gronostaj, 2012) that is based on the English version of FrameNet constructed by the Berkeley research group (Baker et al., 1998). It is found on the SweFN website[1], and is available as a free resource. All lexical resources

[1] http://spraakbanken.gu.se/eng/swefn

used for constructing SweFN are freely available for downloading.

The lexicographers working in this project typically define frames that are fairly close in meaning to their counterparts in the Berkeley FrameNet. When a frame has been defined, lexical units are added. For each lexical unit, a set of example sentences are then selected from KORP, a collection of corpora of different types (Borin et al., 2012). Finally, the lexicographers annotate the frame element (semantic role) structure on the example sentences.

We now proceed to describe the relevance and distance measures used in the FrameNet lexicographic task.

3.1 GDEX-inspired Relevance Measure

As mentioned above, GDEX (Kilgarriff et al., 2004) is a method for extracting example sentences from corpora. The stated purpose is that the selected examples should be

- typical, exhibiting frequent and well-dispersed patterns of usage;
- informative, helping to elucidate the definition;
- intelligible to learners, avoiding complex syntax and rare words.

These goals are of course hard to quantify, but GDEX includes a number of rules of thumb intended to capture these properties. We defined a relevance measure based on a simplified subset of the rules used in GDEX.

Sentence length: if the sentence was shorter than 10 or longer than 25 words, five relevance points were subtracted.

Rare words: one relevance point was subtracted for each infrequent word.

Main clause: since we didn't want to parse the sentence, we just subtracted one relevance point if the search term occurred after the tenth position in the sentece.

3.2 Contextual Distances

To compute distances between the two examples i and j, we used a standard Euclidean distance between feature vector representations of i and j:

$$d(i,j) = \sqrt{\|\phi(i)\|^2 + \|\phi(j)\|^2 - 2\phi(i)\phi(j)}$$

We developed two different feature extraction functions ϕ, based on based on the syntactic and lexical contexts, respectively.

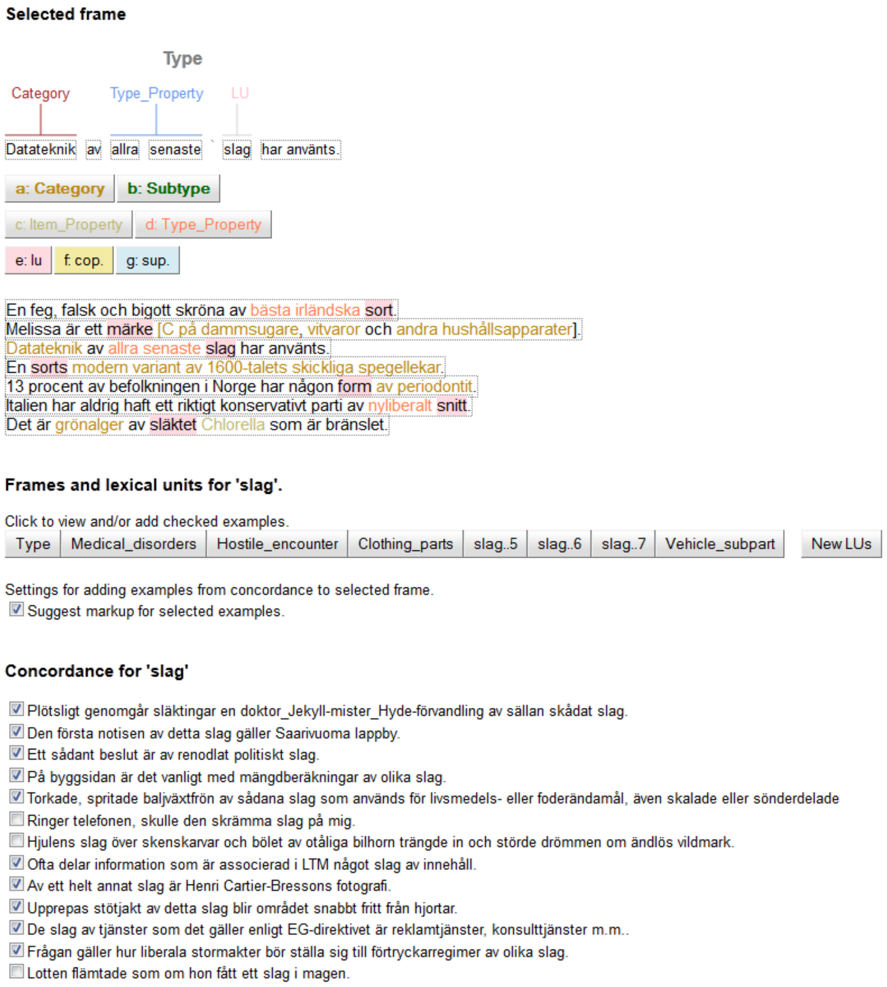
The purpose of the *syntactic* context representation is to distinguish grammatical constructions and subcategorization frames, which is central to the FrameNet lexicographic task. When building the syntactic context representation ϕ_{syn}, we used dependency parse trees provided by MaltParser (Nivre et al., 2007). The trees are pre-computed and stored in the corpus database, so this does not significantly affect the computational performance. The feature vector consists of one feature for each incoming and outgoing dependency relation of each word in the search hit. Direct objects needed some special consideration to take care of reflexives.

The *lexical* context representation uses a standard bag-of-words representation of a window around the search hit. In the future, we aim to compress the feature space by using dimensionality reduction techniques such as random indexing (Kanerva et al., 2000).

3.3 Implementation

Figure 1 shows a screenshot of the user interface for the selection of example sentences for the Swedish FrameNet. The user interface includes an implemenation of the diversification functionality. The implementation proved to be very fast: compared to the time spent iterating through the search result, the diversification added just 14%.

The screenshot shows an example of a diversified result set. We searched for the Swedish word *slag*, and applied the diversification algorithm to produce a set of size 50; we used the GDEX-inspired relevance function and the syntactic context distance measure, and the SUM objective function with a λ of 1. The word *slag* is quite polysemous, with 8 senses listed in the SALDO lexicon (Borin and Forsberg, 2009). In most general Swedish corpora, the completely dominant sense of this word is that corresponding to the English word *type* or *kind*. In the diversified set, we observed 6 of the 8 senses, which shows that the diversification method has worked quite well for this word.



En feg, falsk och bigott skröna av bästa irländska sort.
Melissa är ett märke [C på dammsugare, vitvaror och andra hushållsapparater].
Datateknik av allra senaste slag har använts.
En sorts modern variant av 1600-talets skickliga spegellekar.
13 procent av befolkningen i Norge har någon form av periodontit.
Italien har aldrig haft ett riktigt konservativt parti av nyliberalt snitt.
Det är grönalger av släktet Chlorella som är bränslet.

Frames and lexical units for 'slag'.

Click to view and/or add checked examples.

Settings for adding examples from concordance to selected frame.
☑ Suggest markup for selected examples.

Concordance for 'slag'

☑ Plötsligt genomgår släktingar en doktor_Jekyll-mister_Hyde-förvandling av sällan skådat slag.
☑ Den första notisen av detta slag gäller Saarivuoma lappby.
☑ Ett sådant beslut är av renodlat politiskt slag.
☑ På byggsidan är det vanligt med mängdberäkningar av olika slag.
☑ Torkade, spritade baljväxtfrön av sådana slag som används för livsmedels- eller foderändamål, även skalade eller sönderdelade
□ Ringer telefonen, skulle den skrämma slag på mig.
□ Hjulens slag över skenskarvar och bölet av otåliga bilhorn trängde in och störde drömmen om ändlös vildmark.
☑ Ofta delar information som är associerad i LTM något slag av innehåll.
☑ Av ett helt annat slag är Henri Cartier-Bressons fotografi.
☑ Upprepas stötjakt av detta slag blir området snabbt fritt från hjortar.
☑ De slag av tjänster som det gäller enligt EG-direktivet är reklamtjänster, konsulttjänster m.m..
☑ Frågan gäller hur liberala stormakter bör ställa sig till förtryckarregimer av olika slag.
□ Lotten flämtade som om hon fått ett slag i magen.

Figure 1: Screenshot of the Swedish FrameNet example selection and annotation user interface.

4 Discussion

We have argued that the recent developments in search result diversification in the information retrieval community are relevant for lexicographers. The work described in this paper builds on previous work in two separate communities that we think may benefit from a cross-fertilization. This has not been very common until now; the most related approach is probably that described by de Melo and Weikum (2009), which similarly defined an optimization problem to build a useful set of example sentences. Although similar in spirit to our method, there are some differences: first, our method does not rely on parallel corpora; second, we maintain a clear separation between relevance and diversity.

We see several obvious ways to proceed. The relevance and distance measures described here are our first attempts, and we believe that more sophisticated measures can be devised. Another necessary next step would to carry out an usability and quality evaluation where annotators are asked whether the presence of the diversified set leads to a better overview of usage and a higher quality of the end result. However, the protocol of this type of evaluation is nontrivial to define.

Acknowledgements

The research presented here was supported by the Swedish Research Council (the project *Swedish Framenet++*, VR dnr 2010-6013) and by the University of Gothenburg through its support of the Centre for Language Technology and Språkbanken (the Swedish Language Bank).

References

B. T. Sue Atkins and Michael Rundell. 2008. *Oxford Guide to Practical Lexicography*. The Oxford University Press.

Collin Baker, Charles Fillmore, and John Lowe. 1998. The Berkeley FrameNet project. In *Proc. of Coling/ACL*.

Lars Borin and Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, Odense, Denmark.

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2010. The past meets the present in the Swedish FrameNet++. In *Proc. of EURALEX*.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC-2012 (to appear)*.

Gerard de Melo and Gerhard Weikum. 2009. Extracting sense-disambiguated example sentences from parallel corpora. In *Proceedings of the First Workshop on Definition Extraction in conjunction with RANLP 2009*, pages 40–46, Shoumen, Bulgaria.

Marina Drosou and Evaggelia Pitoura. 2009. Diversity over continuous data. *IEEE Data Eng. Bull.*, 32(4):49–56.

Marina Drosou and Evaggelia Pitoura. 2010. Search result diversification. *SIGMOD Record*, 39(1):41–47.

Karin Friberg Heppin and Maria Toporowska Gronostaj. 2012. The rocky road towards a Swedish FrameNet. In *Proceedings of LREC-2012 (to appear)*.

Sreenivas Gollapudi and Aneesh Sharma. 2009. An axiomatic approach for result diversification. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 381–390, New York, United States.

Jonathon Hare, David Dupplaw, and Paul Lewis. 2009. IAM@ImageCLEFphoto 2009: Experiments on maximising diversity using image features. In *Proceedings of the CLEF 2009 Workshop*, page 42.

Pentti Kanerva, Jan Kristoffersson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd annual conference of the cognitive science society*.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz, and David Tugwell. 2004. The Sketch engine. In *Proceedings of Euralex*, pages 105–116, Lorient, France.

Adam Kilgarriff, Miloš Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII Euralex international congress*.

Ralf Krestel and Nima Dokoohaki. 2011. Diversifying product review rankings: Getting the full picture. In *Web Intelligence*, pages 138–145.

Enrico Minack, Wolf Siberski, and Wolfgang Nejdl. 2011. Incremental diversification for very large sets: a streaming-based approach. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 585–594, New York, NY, USA. ACM.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2).