Finding Negative Key Phrases
for Internet Advertising Campaigns using Wikipedia

Martin Scaiano
University of Ottawa
mscai056@uottawa.ca

Diana Inkpen
University of Ottawa
diana@site.uottawa.com

Abstract

In internet advertising, negative key phrases are used in order to exclude the display of an advertisement to non-target audience. We describe a method for automatically identifying negative key phrases. We use Wikipedia as our sense inventory and as an annotated corpus from which we create context vectors and determine negative phrases, which correlate with negative senses of a positive key phrase.

1 Introduction

Online advertisers select and bid on key phrases for which their ads will be displayed. Each time an ad is displayed, it is called an impression. The cost of the advertising campaign may be tied directly or indirectly to the number of impressions: each impression may cost the customer, or pricing maybe based on the rate of response to an advertisement (click through rate). Either way, impressions to those who are not interested in a product can cost the advertiser.

Key phrase advertising attempts to infer the interest of a "searcher" from the key phrases they are searching for. Many key phrases have multiple meanings; thus extra phrases are required to accurately infer meaning. There are two approaches that can be taken to avoid uninterested "searchers": over-specification to exclude all other meanings of ambiguous key phrases; and explicit exclusion of some expressions, called negative key phrases.

Over-specification would simply involve selecting long enough and specific enough key phrases so that all ambiguity is removed. This can lead to an explosion of key phrases, many of which may never occur. Finding all possible specific key phrases could be an exhaustive task. Managing all these key phrases could become cumbersome;

but worst of all, interested "searchers" who use the ambiguous terms may not see the advertisement.

Negative key phrases refer to phrases that suggest that a "searcher" is not interested in the product. Thus, in the case of an ambiguous query the advertisement should be shown; but if any negative keywords are present in the query, then the advertisement should not be shown. Negative keywords are a basic function of internet advertising platforms such as Adwords by Google.

There are two types of negative keywords: negative emotions and negative meanings. Negative emotion keywords indicate a "searcher" has negative feelings about the product. Negative meaning refers to unrelated alternate senses of a keyword. Consider for example if you were selling Toyota Corollas (a car). You would likely bid on the word "Corolla" in hopes of attracting a customer. Some negative emotional keywords might be "lemon" or "defects", which suggest the "searcher" has negative sentiments regarding the vehicle. A negative sense keyword might be "flower", which strongly suggests the "searcher" is interested in flower petals (corollas) and not the car.

Our goal is to develop an automated method for selecting negative keywords for an advertising campaign. Automated selection of negative keywords would reduce the effort required to set up such the campaign. In this paper we do not address negative emotion words; we focus solely on identifying negative sense keywords for ambiguous key phrases. It should be noted that negative key phrases are frequently single words, but they can be composed of multiple words.

In section 2 we describe the past research which has guided our work. In Section 3 we describe the method for selecting negative keywords. Section 4 is our evaluation of the method using data from our industrial partner. Section 5 presents our conclusions.

2 Background

Our inspiration for identifying negative keywords comes from word sense disambiguation (WSD) literature (Navigli, 2009). Mihalcea (2007) describes a method for using Wikipedia as a sense inventory and as an annotated corpus for word sense disambiguation. Mihalcea considers each Wikipedia article (or page) as a sense and links as annotated examples of that sense. Each link provides an annotated realization of the sense. In this way, Wikipedia may be considered a partially-labelled corpus and a word sense inventory.

It seems natural to consider the text of an article as related or context words, instead we use the text in the paragraphs containing links to an article as contexts. In other words, instead of using the definition and description as context (Lesk, 1986; Pedersen, 2002), the words around realizations of a sense are used as context.

Wikipedia seems a good choice of sense inventory and corpus for our task, because it is a broad resource covering many topics and specialized domain terms; also it is constantly being updated with modern terms and information, thus already being adapted to new potential advertising topics. Our current Wikipedia index has over 10 million senses and about 50 million annotated examples.

Our system is intended for use with Google Adwords. Google defined negative keywords in the following way[1]:

> Negative keywords are a core component of a successful keyword list. Adding a negative keyword to your ad group or campaign means that your ads won't show for searches containing that term. By filtering out unwanted impressions, negative keywords can help you reach the most appropriate prospects, reduce your cost-per-click (CPC), and increase your ROI [Return on Investment].

Wordstream[2] provides an interactive tool for selecting negative keywords; a user interactively selects a few positive and negative keywords which bootstrap the process. A 2010 US patent Application (20100185661) uses a variety of historic campaign performance information to select negative keywords. Our method is fully automatic and is intended to setup new campaigns (no historic information is required).

3 Method

The basic concept behind our negative keyword generation system is to create context vectors for all senses of an ambiguous key phrases, then to identify components of the context vectors which correlate highly with negative senses and poorly with the positive sense. This is not complete WSD, since the concern is only explicitly identifying one sense, while all other senses are grouped together as negative senses.

The basic steps of the algorithm are shown in Figure 1, while sections 3.1-3.5 describe each step in more detail.

The method can be applied to a set of positive key phrases or to a single key phrase; most steps only consider a single key phrase at a time, but step 4 is intended to improve processing of sets of key phrases. When processing a set of key phrases steps 1-3 are executed for all key phrases, and then step 4 uses all the resulting information.

3.1 Finding all the senses for a key phrase

Given a positive key phrase, we find all possible senses (Wikipedia articles). To do this, we find all the links containing the key phrase. Then from those links, we collect all the final destination pages, also accounting for redirected pages. The set of destination pages for the key phrase is considered the set of possible senses; each sense includes a frequency metric, that is the number of links to the page that used the given key phrase. To optimize this step, we created an indexed database table of all the links in Wikipedia. We recommend flattening this table by storing not just the link destinations, but, if the destination page is a redirect, the redirected destination page.

Consider the keyword "Corolla"; imagine that the word "Corolla" appears in links on pages A, B, C, D. The links on pages A, B and C go to the Toyota Corolla article, while the links on page D go to flower petals. Thus for Corolla the possible senses are Toyota Corolla and flower petal, with frequencies of 3 and 1 respectively.

3.2 Generating context vectors for each sense

Our context vectors are generated from all unigrams (though larger n-grams can be considered) in all paragraphs containing links to a possible

---

Figure 1: Algorithm for selecting negative keywords

1. Find all the senses for a key phrase

   (a) Get all pages referred to by links containing the key phrase.

2. Generate context vectors for each sense

   (a) Find all links referring to this sense;
   (b) Create a vector of words appearing in the paragraph containing the referencing link.

3. Identify the intended sense

   (a) If only one sense exists, mark it as the intended sense;
   (b) else use the most frequent sense or context vector comparison to select the intended sense.

4. Create a broad-scope intended-sense list

   (a) Collect all intended senses for a collection of key phrases (usually key phrases are from an ad group or campaign).

5. Find negative key phrases

   (a) Assign tf-idf values to words (components) in the context vector;
   (b) Divide all related senses between two lists: intended and unintended senses;
   (c) Find the words from the context vectors of the unintended senses that have the highest tf-idf and that do not appear in the context vectors of the indented senses.

sense. In other words, for each possible sense we use the database table of all the links to find all the pages referring to a particular sense. We then tokenize each of the paragraphs containing a link to the sense being considered. All the words are recorded and counted as a dimension in the vector.

Continuing our previous example, imagine a Toyota Corolla article also has references on pages X and Y (perhaps the link text is "Toyota small car"); while the flower petal article is referred to on page Z (with link text "flower petals"). We would generate a context vector for Toyota Corolla from pages A, B, C, X, and Y; and a context vector for flower petal from pages D and Z. Generating the context vector simply involves counting the words, in the paragraphs where the links appears.

3.3 Identifying the Intended Sense

There are many ways that the intended sense can be assigned, depending on the resources available. WSD could be applied to an example context if one is available; in our case examples are likely the ads from the advertising campaign.

A simple WSD method that can be used when no examples are available is selecting the most frequent sense of the key phrase; this can be deter-

mined using the frequency information from step 1. We found that this method works quite well when multiple key phrases are being processed because step 4 will compensate for a few mislabeled senses. When examples are available, another simple WSD method is to compare the context vectors of a sense to the example contexts (in our case advertisements) and choose the sense with the most similar vector.

3.4 Creating a broad-scope intended-sense list

This step is only relevant if multiple key phrases are being processed. This step requires all intended senses for all key phrases. We collect all the intended senses of all key phrases into what we call the broad scope intended sense list. There are a number of cases where a key phrase may have more than one intended sense, using this method we collect all the intended senses and avoid blocking secondary intended senses.

False positive senses will generate unwanted impressions, which are undesirable, but false negative senses are more problematic because an ad may not be shown to the intended audience. There are often multiple positive key phrases assigned to

any single sense and thus, by collecting all the intended senses, we reduce the risk of assigning a false negative sense. We observed that, even if a single key phrase is mislabeled (in our case due to choosing the most frequent sense), the correct label was consistently identified by other keywords.

Furthermore, the collection of these senses could be used with clustering or other techniques that might reveal additional senses that should have been considered. These additional senses may even provide new positive key phrases.

Consider setting up an advertising campaign for Toyota Vehicles. A small selection of key phrases that might be used in this campaign is: "Corolla", "Sienna", "Toyota minivan". If each key phrase was assigned the following senses, respectively, then the broad scope intended sense list would be: "Toyota Corolla", "Sienna Miller", "Toyota Sienna". "Sienna Miller" (an actress) is in fact a mislabeled sense, but due to other keywords, the correct sense has been included in the broad scope intended sense list, thus avoiding a false negative.

3.5 Finding negative key phrases

We divide all senses of a positive key phrase into two sets of senses: the positive set (anything in the broad scope sense list), and the negative set (everything else). We evaluate all components of the context vectors from all senses: first we evaluate the components (unigram, bigram, etc.) using tf-idf (Salton, 1989), where *tf* is simply the frequency from Step 1 and *idf* has been pre-calculated from the Wikipedia corpus. We then select the N highest valued (tf-idf) components above a minimum threshold, from the negative set, and then confirm that each component either never appears as a component in the positive set, or that the positive set tf-idf is below a choosen threshold.

4 Evaluation

We used existing campaign data from our industrial partner as test data. We generated lists of negative keywords for the key phrases in an existing ad campaign. We could not consider the existing negative keyword lists from the campaign as a gold standard because they were incomplete, for only a few topics; they also contained intentional misspellings (something that this system does not consider); they contained negative keywords that were related to user intention instead of meaning (such as car rentals instead of purchases); and they

contained a few emotional negative keywords indicating that the audience had negative sentiments towards the focus of the ad.

Thus, we used a number of ad hoc tests to determine the effectiveness of our system. First, we examined the results for any obvious patterns or flaws. We discuss these impressions and considerations in the subjective evaluation. To empirically and more objectively evaluate the effectiveness of our negative keyword selection, we collected metrics from Google's Adwords evaluation tools; these are discussed in the empirical evaluation. Table 1 summarizes both the empirical and subjective evaluation.

4.1 Subjective Evaluation

In table 1, all of the negative keywords for "Toyota Sienna" are strongly associated with some topic other than "Toyota Sienna", "minivans", etc. Most of them refer to people or groups. This is probably the nature of the word "Sienna", which is normally a proper name. It should be noted that "Miller" does not appear in the list, even though it would be an effective negative keyword removing the sense "Sienna Miller". It would in fact rank higher than our current number 1, but due to an error in the overly simple sense disambiguation method "Sienna Miller" was consider a correct sense, along with "Toyota Sienna".

In table 1, the top ten negative keywords for "Corolla" refer almost entirely to a single topic: the "flower petals". "Corolla" is not a common word and thus this was probably one of the only alternative senses. Neither "Corolla petals" nor "corolla flowers" were frequent enough to be listed in the Google traffic estimator. All our searches for these terms produced "Toyota Corolla" advertisements, even though none of the top articles were about Toyota Corollas.

4.2 Empirical Evaluation

We limited our empirical evaluation to the top 10 negative results for 5 positive keywords. Our first empirical evaluation was against the existing negative keywords from the campaign, but only 2 of the 50 negative keywords existed in the list, though a small number of thematic correlations existed. The choice of words were different, but often words with similar senses were present. The generated negative keyword list provided a number of senses not covered under the original campaign list.

We collected the following metrics from the Google Adwords evaluation tool for each pair of positive and negative key phrases:

1. How many of the top ten search results were related (in any way) to a positive sense. We hoped to evaluate whether the key phrases were in fact highly correlated with a positive or negative sense.

2. What was the estimated monthly search frequency for the combined positive and negative key phrase pairs. This would help determine the effectiveness or utility of the negative key phrases.

3. Were the campaign ads (or very closely related ads) shown or not. This would determine whether these key phrases would be beneficial to the campaign.

Positive	Negative	Freq.	Top Ten	Ad
Corolla	petals	0	10	●
Corolla	flowers	260	10	
Corolla	sepals	0	10	●
Corolla	flower	880	10	
Corolla	centimeters	0	9	●
Corolla	stamens	0	10	●
Corolla	species	0	10	
Corolla	flowering	0	10	
Corolla	fruit	0	10	●
Corolla	calochortus	0	10	●
Corolla	erect	0	10	●
Avalon	b0e0e6	0	10	
Avalon	webcomic	58	10	
Avalon	frankie	22200	10	
Avalon	newfoundland	1300	10	
Avalon	ranavalona	0	10	
Avalon	avalonia	0	10	
Avalon	peninsula	1900	10	
Avalon	arthur	2400	10	
Avalon	mists	27100	10	
Avalon	funicello	880	10	
Avalon	laurentia	0	10	

The top ten search results for 44 of the 50 evaluated key phrase pairs were entirely about the negative topics. There were only two cases where at least half of the top ten results were related to a positive sense. This suggests that the system generally provides negative key phrases that are not correlated with the positive senses.

Positive	Negative	Freq.	Top Ten	Ad
Highlander	league	260	10	
Highlander	baseball	0	10	○
Highlander	sox	0	10	○
Highlander	scottish	2900	10	
Highlander	highlands	390	10	
Highlander	pitcher	0	10	●
Highlander	team	320	10	
Highlander	nyy	0	10	●
Highlander	player	0	10	
Highlander	boston	0	2	●
Highlander	dodgers	0	7	
Sienna	guillory	74000	10	
Sienna	edward	0	7	
Sienna	louis	0	8	
Sienna	france	0	10	
Sienna	emperor	0	10	
Sienna	pope	0	10	
Sienna	burnt	8100	10	
Sienna	samuel	0	10	
Sienna	jackson	110	5	
Sienna	bargagagli	0	7/7	○
Sienna	hollzman	0	10	
Tacoma	he	0	10	●
Tacoma	rainiers	14800	10	
Tacoma	soccer	2900	10	
Tacoma	season	0	10	○
Tacoma	airport	18100	10	○
Tacoma	league	0	10	○
Tacoma	bridge	40500	10	
Tacoma	indoor	0	10	●
Tacoma	mariners	0	10	
Tacoma	dome	33100	10	
Tacoma	seattle	40500	10	

Table 1: Empirical evaluation of results.

21 key phrase pairs had an estimated search frequency of over a hundred times a month. 8 of the 50 pairs were estimated to be searched tens of thousands of times each month. The existing campaign could save over a hundred thousand impressions to uninterested costumers using our negative keyword list. This metric also showed that about half of our keywords were either for infrequent topics or just infrequent terms; perhaps the estimated monthly search frequency should somehow be considered in the Step 5 where negative keywords are selected using tf-idf. It should be noted that negative key phrases do not cost the advertiser and thus adding infrequent key phrases is

not harmful.

12 of the 50 key phrase pairs triggered ads from our campaign and 20 of the 50 had ads closely related to our campaign. Thus, there are situations where an unintended audience is shown the ad. Note that there seems to be an inverse relation between the estimated number of searches per month and the presentation of the ads in negative contexts. We believe Google Adwords has already implemented some form of sense disambiguation for frequently-searched key phrases; it seems that frequently-searched negative senses for ads are already filtered out. Even if Google may have such a system in place, the addition of negative key phrases does not cost a campaign, may be of assistance on other advertising platforms, and safeguards against any failure of Google's system.

Table 1 shows a selection of the raw data from our evaluation. The first column indicates the positive keyword for which the negative keyword was generated. The second column is the generated negative keyword. The third column represents the estimated monthly search frequency. The fourth column indicates how many of the the top search results, when searching for the positive and the negative keywords together, were unrelated to the intended positive topic; a result of 10 indicates the results are completely unrelated to the positive topic. Please note that all the positive topics here refer to automobiles from Toyota. The last column is marked with a bullet (•) if a campaign ad was shown, a circle (○) if a related ad was shown, and left empty if all the ads were unrelated to the positive topic. A bullet means impressions are likely given to the wrong audience, while no bullet is ideal to the advertiser.

5 Conclusions

We conclude that our system for negative key phrase generation using Wikipedia effectively finds negative topics, finds words strongly correlated with negative topics, and can improve internet advertising campaigns. Yet we must again state that it seems Google Adwords does not (at least partially) show unintended ads for frequently searched terms.

While we have mentioned that Wikipedia is a broad sense inventory covering many domains, it still has a number of lexical limitations in very specific domains. We observed that with a few very domain specific acronyms and terms (such as names of US government regulations), there was either no appropriate sense or no realization of a particular key phrase.

This paper presented one component of an automated system for configuring internet advertising campaigns. Other components include key phrase extraction and generation for advertisements, grouping (clustering) of keywords and advertisements, and optimization by automated analysis of historic campaign performance.

Further research could include improving Step 4 to identify additional senses through sense clustering. This effort may be combined with keyword generation (selection of non extractive keywords) another component of our industrial partnership.

Future work may include research into better evaluation methods for negative keyword selection. Ironically, evaluation methods may ascribe an estimated value about the effectiveness of a negative key phrase, and thus the evaluation may in turn be a selection method.

Acknowledgments

References

M Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Proceedings of the 5th annual international conference on Systems documentation*, Jan.

R Mihalcea. 2007. Using wikipedia for automatic word sense disambiguation. *Proceedings of NAACL HLT*, Jan.

R Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, Jan.

T Pedersen. 2002. A baseline methodology for word sense disambiguation. *Lecture Notes In Computer Science*, Jan.

G Salton. 1989. Automatic text processing. *Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA ©1988*, Jan.

---

[3]http://www.apption.com