Internet, a *True* Friend of the Translator: The Google Wildcard Operator

Alexander Gelbukh and **Igor Bolshakov**

Center for Computing Research (CIC),
National Polytechnic Institute (IPN),
Mexico City, Mexico
gelbukh@gelbukh.com, igor@cic.ipn.mx; www.Gelbukh.com

We have previously suggested that Internet word usage statistics can help selecting a word choice variant out of a set of hypotheses. In this paper we show how Googe's wildcard operator can help either to directly obtain the specific word appropriate in a given context or to assure that the correct variant be included in the set of considered hypotheses.

Introduction

In the paper we mainly discuss the case of translation from one's native language into a foreign language, or—which is approximately the same task—composing a text in a foreign language, e.g., English. In [2] we have presented a method for guessing the right translation of a word using an Internet search engine. Basically, the method consists in the following:

- Guess several possible translation variants;
- Formulate a query as a hypothetical use of each variant in texts, and perform an Internet search;
- Use the page count returned by the search engine for each query to choose the most frequent variant.

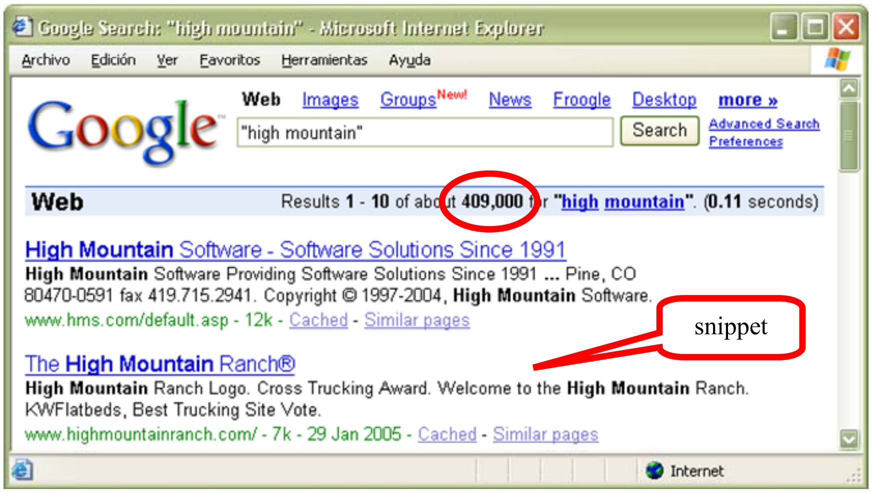
For example, suppose you are not sure whether the Spanish phrase *montaña alta* is translated into English as *tall mountain* or *high mountain*, dictionaries giving both variants for the Spanish word *alto*. A Google search using the two queries gives (all searches in this paper were performed on January 30, 2005 using Google):

Query	Number of documents found
`"tall mountain"`	23,200
`"high mountain"`	409,000

This suggests that *high* is a much better candidate in the given context than *tall*, apparently contrary to one's expectation based on such well-known cases as *tall man*, *tree*, *building*.

The number of documents found is indicated by most search engines; e.g., Google indicates it in the right top part of the window as circled in Figure 1. The query is to be enclosed in quotation marks for the search engine to look for the specific work combination and not for the two words mentioned separately somewhere on the page. A number of additional precautions and potential subtle problems one should take into account are discussed in detail and shown by examples in [2]. Here are some of them:

1. Verify that in the documents or snippets returned by a search engine, the word appears in the intended meaning, syntactic construction, and a document of a suitable style. If in many pages returned for the query this is not the case, re-formulate your query.
2. Restrict your search to the language in question, to avoid coincidence of the words with words in another language.
3. Try to restrict your search to the country where the language in question is spoken. This is especially important for English, which is used by a huge number of non-native speakers.
4. Specifically, avoid (if possible, even by checking the names of the authors of the texts) counting in, or taking example from, pages written by people with the same native language as yours, since they would probably (erroneously) use the same constructions that look "natural" for yourself.



Web Results 1 - 10 of about 409,000 for "high mountain". (0.11 seconds)

High Mountain Software - Software Solutions Since 1991
High Mountain Software Providing Software Solutions Since 1991 ... Pine, CO 80470-0591 fax 419.715.2941. Copyright © 1997-2004, High Mountain Software.
www.hms.com/default.asp - 12k - Cached - Similar pages

The High Mountain Ranch®
High Mountain Ranch Logo. Cross Trucking Award. Welcome to the High Mountain Ranch. KWFlatbeds, Best Trucking Site Vote.
www.highmountainranch.com/ - 7k - 29 Jan 2005 - Cached - Similar pages

snippet

Figure 1. Number of pages found indicated by Google.

5. Do not check only one hypothesis but compare several hypotheses. For nearly every error there are (many) pages on the Internet containing this error, so that a mere existence of a number of pages containing an expression does not indicate that such an expression is correct.

As one can see, to apply the method suggested in [2] it is crucial that the correct variant of the word be included in the set of hypotheses. The method only helps to choose one of the formulated hypotheses but does not provide a clue on how inclusion of the correct variant in the hypotheses set can be assured. This is especially important in light of the penultimate precaution above: people with the same native language as yours will likely use the same construction as looks "natural" for you, which will misleadingly confirm one of the variants you consider given that you did not consider at all the true variant. In the next section, we will give a corresponding example and show that Google's wildcard operator can remedy the situation by helping to include the correct variant in the set of considered hypotheses.

Google's wildcard operator

To choose a preposition in the phrase *He presented an interesting talk _?_ the specialized conference in Honolulu*, one has to guess the possible variants, and then check the statistics for each one through the Internet. The preposition *na* used in the native language of the authors—which is Russian—is literally translated into English as *on*, so Russian native speakers tend to express this idea in English as **talk on the conference*. The preposition *en* used in the second language that we know best—which is Spanish—is literally translated into English as *in*, which gives **talk in the conference* used by many Spanish native speakers. Comparison of the two hypotheses gives:

Query	Number of documents found
`"talk on the conference"`	133
`"talk in the conference"`	323

which apparently selects **talk in the conference*. However, there are two warnings in the search results: first, the low absolute numbers of occurrences and second, the countries where the pages are found: the top-level domains of the first several pages returned for this query either belong to India, China, France, Japan, Spain, Zambia, Sweden, Denmark, Taiwan, Germany, or are language-neutral, such as .com, .edu, or .net.

The same situation holds for the phrase *He returned home _?_ taxi*: Russian and Spanish use the same prepositions as in the previous example. Adding several more "plausible" variants (our intuition giving no beforehand clue on their plausibility), we get the following results:

Query	Number of documents found
`"go on taxi"`	29
`"go in taxi"`	54
`"go with taxi"`	213
`"go inside taxi"`	6
`"go at taxi"`	1

In both cases, the right hypothesis cannot be formed (by a native speaker of Russian or Spanish) without knowing the right answer (or without very good language intuition: the latter table does give an idea of "instrumental" role of *taxi*).

One can try searching for the two words without indicating the word in between. Obviously, just omitting the word and keeping the quotation marks in the query does not give the desired results. However, omitting quotation marks (and thus allowing the two words to appear in any place on the page in the search results) gives too much garbage:

Query	Extracts from the obtained snippets
`"go taxi"`	**Go TAXI** verbindet provisionsfrei; Trains **GOTaxi** GOWater; Pick-N-**Go Taxi**; **Go Taxi** Go!; where they **go**. **Taxi** is not expensive
`go taxi`	It's A Theater In A **Taxi**...It's An Internet Connected Laser Light Show Webmobile In **...** TaxiCam Photo 1/15/2005 Jason And Daniel **Go** To The Fireworks; if **taxi** operators **go** ahead; SVO and **taxi** in Moscow. **... GO**-TO.RU.

A solution for the problem is Google's wildcard operator "*", which thus can help guessing the right word without knowing it in advance. In Google's queries, an asterisk can be used as a placeholder of one word, without specifying it. More than one asterisk can appear in the query, which gives two or more unspecified words between the specified ones. Here are extracts from the first 10 Google snippets on the following two queries (recall that the quotation marks in the query are essential here):

Query	Extracts from first 10 snippets
`"talk * the conference"`	to **talk about the conference** agenda; to present a **talk at the conference**; for Zaninotto's **talk**; **the conference** dinner featured; we'll **talk about the conference**; to **talk up the conference**; Invited **talk at the conference**; **Talk about the conference**; TOUGH **TALK AT THE CONFERENCE** BOARD; **talk**, **for** > **the conference** delegates; we'll **talk about the conference**
`"go * taxi"`	10 minutes to **go by taxi**; can **go by taxi**; should i just **go by taxi**; IN-**GO searchstring**: **taxi** driver; then **go by taxi**; wouldn`t **go by taxi**; G-**GO taxi** S1

Excluding the expressions with a clearly different meaning than the intended one, we immediately get the answer. Comparison of the newly found variants with the ones considered previously shows that they are clear winners:

Query	Number of documents found
`"talk at the conference"`	8600
`"go by taxi"`	4370

Naturally, this technique can be applied not only to find the prepositions but also content words, if you can guess a right context. For example, the situation expressed in mathematics by a formula $x \to \infty$, in plain Russian is described with the verb *stremitsya* (*k* 'to'), for which the dictionary [4] gives English translations *speed*, *rush*, *seek*, *aim*, *aspire*, *strive*, *long*, *crave*: *The variable x ? to infinity*. A Google search gives the following result:

Query	Extracts from first 10 snippets
"x * to infinity"	If f(**x**) **tends to infinity** as x tends to c, then; as **x goes to infinity**; when **x tends to infinity**; is **x going to infinity**; if **x goes to - infinity**; as **x goes to infinity**; as **x goes to infinity**; as **x tends to infinity**; as **x tends to infinity**; as **x tends to infinity**

A comparison of the obtained candidates (including literal translations from Russian):

Query	Number of documents found
"x speeds to infinity"	0
"x rushes to infinity"	0
"x strives to infinity"	0
"x goes to infinity"	618
"x tends to infinity"	501

suggests that the latter two variants (none of which is present in the dictionary) are nearly equally used, probably *tends* being a more formal and *goes* a more colloquial one.

Finally, a clarification is appropriate here: as has been pointed out in [2], Google does not count correctly the number of pages, at least for non-standard queries [1, 3]. This can be illustrated with the following results:

Query	Number of documents found
"talk at the conference"	8600
"talk at a conference"	908
"talk at * conference"	3520
"talk * the conference"	2900
"talk * * conference"	949

Yet another inconsistency which we observed as to the use of the wildcard operator is that Google seems to sometimes interpret it as “one word or nothing” instead of “exactly one word”. Though we cannot comment on the reasons for such inconsistencies, they seem not to present any serious problems for the suggested technique.

Conclusions and future work

In an earlier work [2] we have presented a method of verification of word choice variants through Internet search engine statistics: one forms several possible hypotheses and chooses the one that is most used on Internet. However, in that paper we did not consider how one can guess the possible variants without knowing them in advance, i.e., how one can assure the true variant is on the list of hypotheses under comparison. Here we have shown how Google’s wildcard operator “*” can greatly help in finding the correct word in a given context.

The suggested technique can be rather straightforwardly automated, leading to the development of a useful tool for computer-aided translation and text authoring. However, a number of subtle problems, such as word sense ambiguity and comparison of the word usage observed on the found Internet pages with the intended meaning of the translation at hand, currently makes totally automatic use of the described technique in machine translation and automatic text generation applications non-trivial.

Acknowledgement

Work was done under partial support of Mexican Government (SNI, CONACyT, CGPI-IPN).

References

1. Bolshakov, I. A., S. N. Galicia-Haro. Can We Correctly Estimate the Total Number of Pages in Google for a Specific Language? In: A. Gelbukh (Ed.). *Computational Linguistics and Intelligent Text Processing*. Proc. 4th Intern. Conf. on Computational Linguistics, CICLing-2003. *Lecture Notes in Computer Science* 2588:415–419, Springer-Verlag, 2003.

2. Gelbukh, A., I. A. Bolshakov. Internet, a true friend of translator. *International J. of Translation*, 15(2): 31–50, 2003; www.Gelbukh.com/CV/Publications/2003/Bahri-2003.htm.

3. Notess, G. R. Google Inconsistencies. www.searchengineshowdown.com/features/google/inconsistent.shtml#count.

4. Smirnitsky A. I. *et al*. Large Russian-English Dictionary (in Russian). Russkij Jazyk, 2004, pp. 768.