Towards Analysis Tools for a Multilingual Blogsphere

Douglas W. Oard

College of Information Studies and Institute for Advanced Computer Studies
University of Maryland, College Park, MD, 20742
oard@umd.edu

Abstract

Many approaches to Weblog analysis draw on words found in the postings and comments, and the meaning of those words is inherently bound to a language. Moving beyond single-language analysis will require tools that leverage translation resources in ways that are well matched to the task. In this paper, we describe how state of the art techniques developed originally for machine translation and cross-language information retrieval could provide a foundation for extending existing blog analysis tools to learn from many languages things would be harder to see in just one.

Introduction

A 2003 survey of about a half a million blogs reported that about one third (36%) of the surveyed blogs did not appear (to automatic language identification software) to be written principally in English.[1] Some blog search engines (e.g., bloogz.com) allow the user to select their desired search language, and one service (portal.eatonweb.com) lists 49 known blog languages. With the number of blogs still doubling every 5 months or so, clearly there is a lot of multilingual content out there.

So long as the task involves looking at content in only one language, techniques that work well in some language can generally be adapted fairly easily to handle another. For example, BlogPulse's tracker shows that "espana" occurs far more often in than "estados unidos," the reverse of what we would see when those Spanish terms are expressed instead in English. But tasks such as multi-perspective summarization that would require simultaneous analysis of content in multiple languages are harder to support—for that, we want tools that leverage the available resources for translating meaning between languages.

Over the past decade, term-based techniques for crossing the language barrier have been extensively investigated for information retrieval applications. Although that is a fairly

[1] http://www.blogcensus.net/

specialized task, it turns out that the cross-language mapping component is both quite modular (and thus easily reused) and remarkably effective (and thus worth reusing). The next section briefly reviews the key ideas for applying a translation mapping to a term-based analysis problem. The two sections that follow then describe two approaches to mapping the results of multilingual analysis into a single language. The paper concludes with some suggestions for next steps.

Term-Scale Meaning Matching

The operation of so-called "vector space" information retrieval systems can be described at a high level as follows:

- The set of terms used in each document are noted, and additional terms that might could have been used to express the same meaning (but weren't) are added. Similar processing can also be done for the query. In monolingual applications, this process is called expansion. In cross-language applications, we call it translation.

- For each term in each document, weights are pre-computed that reflect the contribution of that term to describing the document. These weights typically include factors that reflect the degree to which the term is associated with the document (often called "length-normalized term frequency"), the descriptive power of the term across the collection (often referred to as "inverse document frequency"), and whatever side information is available about the prior probability that a document will be relevant (e.g., computed using centrality in a social network). The actual function used to compute the weights is generally tuned using test collections that are representative of the ultimate application.

- For each query-document pair, a score is computed that describes the degree to which the document addresses the topic specified in the query. A similar process can be used to compute document-document similarity for use in clustering,

classification, and pre-filtering stages of other applications.

Obviously, the translation process is the key to any cross-language application. Importantly, this is not translation of running text, but rather translation of individual terms. Even more importantly, this is not "one-best" winner-take-all translation, but rather translation that admits the possibility of partial counts (Darwish and Oard, 2003). For example, if the system cannot determine whether "happy" is an emotion or a dwarf, both possibilities can be accommodated, and the term occurrence count can be split across the alternatives in a manner that reflects the relative predominance of those two senses observed in a large collection of human-prepared translations. Indeed, the ability to learn from large collections of translations turns out to be the key to effective term-scale meaning matching. The next section looks at that problem.

Blog Translation

Broadly speaking (and at some risk of oversimplification), there are two ways to approach the problem of machine translation. The first relies on hiring linguists to build what amount to sophisticated expert systems. The second relies on building machines that perform statistical analysis of existing translations that were created by people. The two approaches presently yield translations that are about equally good (or bad!), but the statistical machine translation approach is far less expensive, and it can be extended to additional language pairs much more expeditiously. Moreover, the fully automated metrics that are now widely used for formative evaluation of translation quality facilitate iterative refinement of statistical machine translation systems.

Figure 1 shows some results from using Hiero, a state of the art statistical machine translation system (Chiang et al, 2005), to translate a Chinese blog posting into English. While the translated text is quite hard to read, the general topic is easily identified. If we ignore the variant endings (i.e., after stemming) and omit extremely common words (those that account for at least one word in a thousand in typical English), the words in Figure 1 with the highest term frequency are: teacher (6), interview (4), education (3), word (3), Huo (2), Yong (2), full (2), feel (2). More sophisticated weighting functions can yield further improvement (e.g., if "feel" were to occur more commonly in blog postings than in general English). The bottom line is that we can already perform any type of analysis that relies on counting English terms, even with blogs written in Chinese, thus facilitating comparative analysis between English and Chinese. Moreover, achieving similar results for any pair of widely used languages is already well within the state of the art. Some types of analysis depend on features of text other than the mere presence or absence of terms, however, and the disfluencies evident in Figure 1 could clearly be problematic in such cases. That is the focus of the next section.

Teachers and source of data on the Huo Yong interview after Kwun 01 04, 2005
Occasionally took a look at the central bank on columns, well-known and special-grade teacher Huo Yong interview. Sometimes I feel that, on education, we words has practically used it, too blandishments, full of paper, full of media and the politically correct, also fundable language infectious. Interview mentioned no love for such without education, and the poor students, as long as the good teachers. In fact are not considered a fresh words, it is estimated that China's teachers and parents have realized that a penetrate incomparable. If did not see action, who is susceptible to moved. But during the interview, or a word that I feel shock: talked about a children's work, he said the instructor learning, he was doing battle cry. Don't know teachers need from the hearts of owns understanding? Our education and are doing what, and obstruct the teachers in this sense

Figure 1. Hiero translation of a Chinese Blog.

Annotation Projection

Many language processing tasks depend on both the terms that are present and the order in which those terms appear. Examples include sentiment detection (e.g., "I was not happy about that" vs. "I was very happy about that"), recognition of named entities (e.g., "President Bush" vs. "Bush League"), and co-reference resolution (e.g., "Jim criticized Mark, but before he knew it Mark had denounced him"). Presently, we can only automate these types of processing with any degree of accuracy if we do that processing before we do translation. That naturally leads to the question of how we can merge results from multiple cases of language-specific processing to perform cross-language analysis.

One way to do this is to build systems that translate annotations while they translate the text. This turns out to be fairly straightforward because statistical machine translation systems create term alignments as a byproduct of the translation process. Any annotations associated with a set of terms in one language are therefore easily projected to the other language. Of course, it could be expensive to create similar annotation capabilities in many languages. Fortunately, we can apply the same trick one more time for annotation systems built using machine learning techniques. The basic idea is to annotate training material in one language, project those annotations to a second language, and then (usually after a few hand-coded tweaks) use those projected annotations to train an annotation system for the second language (Hwa et al, 2005; Yarowsky and Ngai, 2001).

Next Steps

Much of what we will need to support multilingual blog analysis could be built today, but there are a few issues that need to be addressed before we will be ready to build complete systems. At the top of our list are:

- Better statistical machine translation for blogs. This produces two types of dividends: (1) better alignments, which yield more accurate annotation projection; and (2) better translation probabilities, which yield better term statistics. Data-driven techniques require representative data, and in this case that means that we need some Weblog translations. The system that produced the results in Figure 1 was trained for news stories, and it does much better on news than it does on blogs. Nowson et al found that blog postings used somewhat more contextually bound language than newspaper articles, so larger quantities of training data may be needed to attain comparable translation quality. We know of one blog that is being routinely hand-translated,[1] and the Linguistic Data Consortium has plans to translate 400,000 words of Arabic and Chinese blogs into English.[2] If a worldwide community of interest emerges, similar efforts could be started for additional language pairs.

- Merging term-based analysis results. While mapping term statistics from one language to another is now quite tractable, we know far less about normalizing those statistics in ways that would facilitate comparisons between sources from different languages. This problem has been investigated in the context of cross-language information retrieval, but at present the best known techniques rely on post-combining the results of separate analysis rather than performing the analysis on pre-combined statistics (Braschler, 2004). The fundamental problem is that we do not yet have good models for the types of errors that our translation systems make (Diekema 2003). Focusing on new analysis tasks may, however, reveal new opportunities.

The ideas in this brief paper are intended as a starting point for a conversation about how they could be used with blog analysis techniques that now exist in a monolingual context or that could be built. Of course, we need to start that conversation with a clear picture of the kinds of capabilities that are needed. With technorati.com now reporting more than 20 million blogs, there should be no shortage of data or potential users!

---

[1] afamilyinbaghdad.blogspot.com posts the same content in Arabic and English.
[2] http://projects.ldc.upenn.edu/gale/data/DataMatrix.html

References

Martin Braschler, *Robust Multilingual Information Retrieval*, Ph.D. Dissertation, University of Neuchatel, Switzerland, 2004.

David Chiang, Adam Lopez, Nitin Madnani, Christof Monz, Philip Resnik, and Michael Subotin, "The Hiero Machine Translation System: Extensions, Evaluation, and Analysis," *Proceedings of HLT/EMNLP 2005*, Vancouver, October 2005.

Kareem Darwish and Douglas W. Oard, "Probabilistic Structured Query Methods," *Proceedings of ACM SIGIR 2003*, Toronto, July 2003.

Anne R. Diekema, *Translation Events in Cross-Language Information Retrieval: Lexical ambiguity, lexical holes, vocabulary mismatch, and correct translations*, Ph.D. Dissertation, Syracuse University, 2003.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas and Okan Kolak, "Bootstrapping Parsers via Syntactic Projection Across Parallel Texts," *Natural Language Engineering*, June, 2005.

Scott Nowson, Jon Oberlander, Alastair Gill, "Weblogs, Genres, and Individual Differences," *Proceedings of the Cognitive Science Society*, Stresa, Italy, 2005.

David Yarowsky and Grace Ngai, "Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection Across Aligned Corpora," *Proceedings of NAACL-2001*, Pittsburgh, June, 2001.