Twasebook: A "Crowdsourced Phrasebook" for Language Learners using Twitter

Graeme W. Coleman, Nick A. Hine
School of Computing
University of Dundee
Dundee, Scotland, UK, DD1 4HN
{gcoleman, nhine}@computing.dundee.ac.uk

ABSTRACT

We present Twasebook, a web application aimed at learners of foreign languages. Based on a learner's search terms, Twasebook identifies relevant status updates from Twitter in the learner's target language(s) as examples of everyday vocabulary. Twasebook therefore represents a proof of concept application designed to explore the feasibility of utilizing the vast amounts of open content generated by social networking tools within the context of language learning. In this paper, we present the motivation behind Twasebook, a brief description of how the application works, and our plans for future development.

Author Keywords

Twitter; language learning; social networking; microblogging; crowdsourcing.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Human Factors; Design.

INTRODUCTION

Since launching in 2006, the Twitter microblogging service has quickly become a cultural phenomenon. Recent figures suggest that around a quarter of a billion *tweets* (status updates by individual users) are posted every day [4]. When a major international event occurs, such as a royal wedding or the death of a celebrity, there can be between four to eight thousand individual tweets commenting on the event every second [5].

A particularly interesting aspect of Twitter for language learners is that there is no single dominant language; one study notes that the top five languages used are English (50%), Japanese (14%), Portuguese (9%), Malay (6%) and Spanish (4%) [3]. Additionally, the Twitter website (http://www.twitter.com) is now available in 30 languages [6], thus opening the concept of microblogging on Twitter up to users who, by being able to use the interface and post tweets in their own language, would otherwise be unable to engage with the application. Consequently, the number of tweets in those "new" languages looks likely to rise further. We suggest that the user-generated content available openly through Twitter, and other social networking services, represents a potentially fruitful vocabulary corpus for learners of foreign languages seeking to increase their exposure to their chosen target language(s) (particularly the idiomatic flavors which are often not touched upon in more formal language textbooks), furthering their understanding of how the language is used by native speakers in their everyday lives, and possibly even encouraging the curious amongst us to begin the process of engaging with a new language.

TWASEBOOK

Twasebook is a web application built upon the Twitter Search and Microsoft Translator APIs (a beta version of the application is available at http://www.twasebook.org). The pedagogical inspiration for this work is an eLearning technique known as *microlearning*, in which a difficult learning task (such as learning a language) is broken down into a series of manageable chunks distributed over time [2]. This approach is not a replacement for a programme of in-depth learning; rather, it allows learners to chip away at a larger learning goal during idle periods, such as waiting for a bus [1].

Overview

In Twasebook, the learner inserts a word or short phrase in his or her own language (currently English is the default "home" language) and, through the Microsoft Translator API, Twasebook translates this phrase into the learner's target language (each of the 30+ languages included as part of the Microsoft Translator API are supported). Subsequently, Twasebook searches through the last 6-9 days worth of tweets, as restricted by the Twitter Search API and, assuming the translated version of the word/phrase appears within the search results, presents a list of tweets containing this text (as well as a translation of the entire tweet into English). Hence, the learner is provided not only with a direct translation of the original word or phrase, but also examples showing how this text is actually used in an everyday context. We therefore refer to

Twasebook as a "crowdsourced phrasebook", as the examples of use are entirely generated by Twitter users.
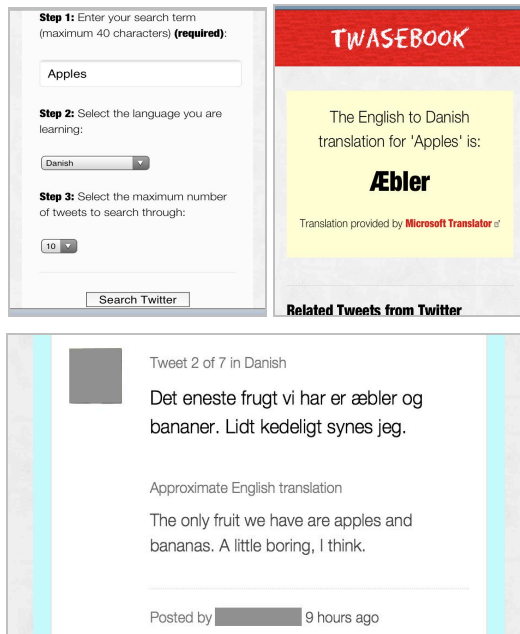

Figure 1. Translation of the English "apples" to the Danish "æbler" alongside a usage example in Danish from a Twitter user, with an approximate translation to English generated by the Microsoft Translator API

Challenges

Aside from technical issues in which the fetching and subsequent translation of results can often be quite slow depending on traffic, the greatest challenges we have encountered thus far relate to the quantity and quality of the results returned. To begin with, there is no guarantee that any Twitter users will have used a particular word or phrase in the relatively short window of tweets that Twasebook searches through. Therefore, uncommon words or phrases (which may be of more interest to language learners) often yield far fewer, if any, results from Twitter than single, common, words. Additionally, words with multiple meanings can result in the "wrong" translation being returned; for example, the word "run" in English has many different meanings depending on the context (e.g. "to run a company", "to run for one's life", "to run a flag up a pole") and therefore the application may return a large amount of examples of use unrelated to the context that the learner expects. Coupled with this, unlike search engines for Web pages, Twitter's search facility lacks relevant measures for ranking content. Instead, Twitter typically returns matches ordered by *recentness* rather than by *relevance* [7] thus there is as yet no way to rank or filter out pedagogically "good" examples of vocabulary usage from tweets containing misspellings or poor grammar which may confuse or be unhelpful to learners.

FUTURE WORK

We are currently developing additional functionality within Twasebook to further assist language learners, such as the ability to rank and store results for later reference, allow the learner to add "notes-to-self" on, for example, interesting grammatical aspects within individual tweets, an individual user-generated categorizing facility to allow learners to group tweets by subject (for example, examples of irregular verb usage, tweets by theme, and so on), and the ability to store "corrected" versions of the translated tweets into the learner's own language (given that computer generated translations such as those generated by Microsoft Translator are not always accurate or grammatically correct). We are also carrying out a series of formal user evaluations with language learners, combining quantitative approaches (i.e. investigating the type and number of searches that participants undertake) and qualitative approaches, in which participants will be asked to use the application over a period of time as part of their learning and to provide feedback on the acceptability and perceived value of the application within a wider language learning context.

ACKNOWLEDGMENTS

REFERENCES

1. Beaudin, J.S., Intille, S.S., Tapia, E.M., Rockinson, R. and Morris, M.E. Context-sensitive microlearning of foreign language vocabulary on a mobile device. In *Proc. European Conference on Ambient Intelligence*, Spring-Verlag Berlin (2007), 55-72.

2. Gassler, G., Hug, T., and Glahn, C. Integrated Micro Learning – an outline of the basic method and first results. In *Proc. Interactive Computer Aided Learning*, Kassel University Press (2004), 1-7.

3. Semiocast, Half of messages on Twitter are not in English. http://semiocast.com/downloads/Semiocast_Half_of_me ssages_on_Twitter_are_not_in_English_20100224.pdf (2010).

4. Tsotsis, A. Twitter is at 250 million tweets per day, iOS 5 integration made signups increase 3x. *Techcrunch* (2011), http://techcrunch.com/2011/10/17/twitter-is-at-250-million-tweets-per-day/.

5. Twitter, Tweets per second. http://yearinreview.twitter.com/en/tps.html (2011).

6. Twitter, Twitter now in Catalan and Ukrainian. http://blog.twitter.com/2012/07/twitter-now-in-catalan-and-ukrainian.html (2012).

7. Van Kleek, M., Smith, D.A., Stranders, R. and Schraefel, M.C. Twiage: A game for finding good advice on Twitter. In *Proc. CHI 2012*, ACM Press (2012), 889-898.