KEYSTONE COST ACTION IC1302, STSM SCIENTIFIC REPORT

STSM TOPIC: Interactive Language Learning: Recommending keywords with POS tagged multilingual word graphs

STSM Applicant: Benjamin Bergner, Otto-von-Guericke-University Magdeburg, Germany

Host: Prof. Dr. Alexandre Miguel Pinto, University of Coimbra, Portugal

Period: 14/03/2016 – 28/03/2016

1. Purpose of the STSM

Learning a new language affords the learner to patiently collect and repeat vocabulary as well as to understand its use within sentences. Looking at interactive language learning software, generally a limited amount of vocabulary is provided that is not sufficient to completely grasp the active and passive word pool (e.g. to have a meaningful conversation or to be able to completely understand foreign literature). Text from web articles as basis for interactive learning seems to be a proper source to fill this gap and to give insights into language structure. By using this type of source for learning, a wide variety of features that rely on text processing based on keywords can be implemented to aid users. The main purpose of the STSM is to illustrate how keyword exploration can foster the user experience in a language learning environment by building up a first tool.

2. Main Results

The prototype site can be accessed at *vidavoca.com*. Even though a considerable part of the time has been put into the engineering and user experience part, this report will purely address how vocabulary can act as a keyword. The general process at the site independent of the provided keyword based features is the following:

1. User chooses a target language he wants to learn
2. One of several domains/categories of interest (e.g. technology, news, etc.) can be chosen
3. An article out of a collection within the domain can be selected by inspecting a summary
4. Displaying the full article as foundation for keyword based features

The reason for using such an approach besides the higher amount of vocabulary is also to transform the daily internet reading habit into another language, making it easier to stick to learning. Within our first conversations, we came up with a usage scenario that enhances the above process. For the advanced learner, it would foster him to inspect the meaning of single foreign but still unknown words occurring in the selected article. One way to do this is to give the reader an own language example where the translation of the selected foreign word is highlighted (see figures 1 & 2). We examined the following criteria that need to be fulfilled to provide such a benefitting meaning exploration tool in order to aid learners to understand foreign texts.

Criterion 1: POS tags of selected foreign words and their translation inside the given example should be equal. Different languages have different POS tags that need to be transformed to a reduced universal tag set. While the German tagger has been obtained by the *textblob-de* library, a bigram tagger has been used based on the *Floresta* tagged corpora provided by *nltk*. Since POS tags are obtained with machine learning techniques and hence can never be perfectly accurate, this is only a soft criterion within the example selection computation.

Criterion 2: The foreign article and example text should have a similar topic in order to create a smooth switch between the articles as convenience for the learner and to increase the probability of having the same meaning. Therefore, cosine similarity with nouns as best meaning provider is used to compare the foreign article with own language articles that already have been parsed. The decision which example will be recommended is based on the following algorithm:

Input: foreign language article fla, set of own language articles OLA, selected foreign word wf, translated word wt, threshold ts ∈ {0,1}
Output: recommended article

1. Calculate cosine similarity of fla and ola for ola ∈ OLA
2. Get POS tags of wf and wt within fla and ola in OLA
3. Get all ola where difference to highest cosine similarity scoring ola is lower or equal than ts
4. If at least one ola remains where POS tags of wf and wt are equal, sort out all ola where POS tags are not equal
5. Return ola with highest cosine similarity score

The threshold hereby can be set between being very restrictive in two directions. Setting it to 0 results in ignoring POS tags completely, setting it to 1 therefore sorts out suggestions with unequal POS tags before considering topic similarity. The algorithm is chosen in this way so that one can individually implement it depending on the accuracy of the used POS taggers and the individual preferences of the significance of topic similarity and POS tag equality. We think that in this particular application, the importance of having the same topic is higher than the correct POS tag because of POS inaccuracies and its behaviour to classify binary. In order to provide similar topics, preferably a high number of articles need to be available and reviewed per foreign word. By updating (replacing older articles with new ones) through news articles, chances increase that international editors provide similar content (e.g. sport news, announcement of new products, international politics).

Figures 1 and 2 are showing a scenario of recommending a German (acts as own language) sentence with the word in question (*verfügbar*) highlighted after clicking at its Portuguese (acts as foreign language) counterpart word *Disponível*. Both have a rectangle around them to be better recognizable. The German example sentence is about the entity *Apple* and an opened beta version of an OS that can be downloaded after registration. The source article in figure 2 has some exemplary, manually inserted

underlined words that may seem to fit the context with words like *app, iPhone, program, click, new, update* and *smartphone*. The green surroundings indicate which words can generate an example.

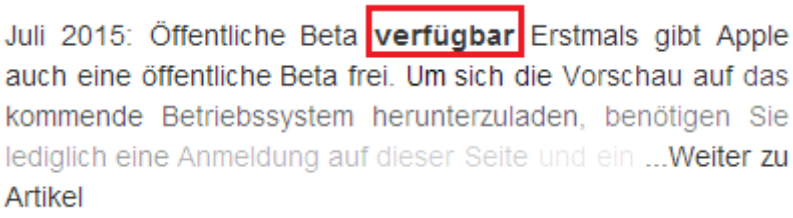

Figure 1: German example sentence given with highlighted translated word

Figure 2: Portuguese (foreign article) with green marked words that can generate an example sentence

Criterion 3: The meanings of foreign word and its translation should be equal given the context due to the translation's ambiguity. This is the strongest and hardest to achieve criterion. Therefore, we came up with the idea to apply English as intermediary language to utilize word definitions and examples of *wordNet* for comparing differences of similarities between those and foreign as well as own language articles respectively. This procedure does not ensure differentiating meaning on a low-level basis considering the variable use cases of a word. The focus here is to detect high-level differences like completely diverse translations (e.g. Portuguese *buscar* can mean *search* or *pick sb. up*).

3. Publications & Future Collaboration

The website can be seen as a playground for implementing keyword based features that are worth to research and in this way worth to use. In occasions of new implementations, we talk about the mechanisms and writing papers. Regarding criterion 3, we stay in touch to formalize and improve the algorithm as soon as different translations per word have been obtained. Besides, I could imagine to do a PhD program in the field of NLU in cooperation with Prof. Pinto.