

Lexical similarity metrics for vocabulary learning modeling in Computer-Assisted Language Learning (CALL)

Ismael ÁVILA and Ricardo GUDWIN
University of Campinas

Introduction

- The L1 can create a basis for learning the vocabulary of an L2: the L1 lexicon helps the learner to infer the meanings of words in L2
- Techniques to compare the word-level distance between L1 and L2 are necessary to model this cross-linguistic influence (incl. quantitatively)

Introduction

- With this metric an ITS can anticipate which L2 words are more easily learned due to transfers from L1 and which ones produce interferences
- The ITS can use this metric to initialize the LM or to sequence the lexical units in terms of their easiness to a particular L1-audience

Introduction

- We present here a technique for measuring lexical similarity in terms of its effect on the learners' perceptual ability in recognizing L2 words with the help of L1 lexicon

Lexical similarity

- Lexical similarities may be due to:
 - Common origin: e.g. Spanish "corazón" and Portuguese "coração"
 - Borrowings: e.g. Japanese "arigato" and Portuguese "obrigado"
 - Coincidences: e.g. Greek "oikia" and Tupi "oca"
- Regardless of their origins, these similarities affect the language learning process and have to be considered by the ITS

Lexical similarity

- The similarity level has two main parallel dimensions: orthographic and phonetic. Each of them may vary from a level of "no similarity" to a level of "absolute match".
- Direction (en) ↔ Direction (fr)
- House (en) ↔ Haus (de)
- Casa (it) ↔ Casa (pt)

Methods to measure string distance

- *Levenshtein distance* uses the minimum number of insertions, deletions and letter substitutions to transform one string into another:

$$LD(s_1, s_2) = \min (n_{ins} + n_{del} + n_{subst})$$

- *Feature distance* is given by the number of features (usually N-grams, substrings of N consecutive letters) in which two strings differ:

$$FD(s_1, s_2) = \max (N_1 + N_2) - m(s_1 + s_2)$$

Where: N_1 and N_2 are the number of N-grams in s_1 and s_2
and $m(s_1 + s_2)$ is the number of matching N-grams

Methods to measure string distance

- The *Levenshtein distance* leads to slightly better classification accuracy but the *Feature distance* allows for much faster searching.
- To account for the fact that one letter change is more relevant in short words than in long ones, normalized versions of LD have been used.

Lexical similarity & language proximity

- An automated method avoids the subjectivity that is inherent in human-made comparisons:

e.g. Gala (el) \leftrightarrow Leche (es)

- We want to measure effective similarity, not linguistic kinship, for similarity, even accidental, is what matters for learning easiness.

Lexical Similarity: perceptual aspects

- A written or printed word is a visual stimulus in the first place.
- Word recognition is easier after fixation of the leftmost than the rightmost letter of a word (the initial in many languages).
- Fixation on the leftmost letter makes the whole word fall in the right visual half-field, in direct connection to the dominant left hemisphere.

Lexical Similarity: perceptual aspects

- Word processing accuracy and speed depend on two factors:
 - Perceptibility of the individual letters as a function of the fixation location
 - The extent to which the most visible letters isolate the target word from its competitors
- The leftmost letters have a special role in word recognition (isolation from competitors).
- Reading and word recognition are not simply based on orthographic information, but involve the activation of phonological codes.

Lexical Similarity: semiotic aspects

- Intuitive word recognition factors are used as a common sense technique when we create abbreviations: *tk*s (*thanks*), *pg* (*page*), *cmd* (*command*) or *ctrl* (*control*).
- Matching initials and consonants is more likely to enable word recognition than matching the same number (same LD) of other letters without the initial or with vowels included: (resp. *tak*, *ae*, *oma*, *coto*).

Lexical Similarity: semiotic aspects

- The recognition of an L2 word due to a similarity with correlated L1 words is an inference based on diagrammatic (iconic) features.
- This “intersymbolic iconicity” explains all the recognitions based on similarity, regardless of their cause: common origin, borrowings or simple coincidence.

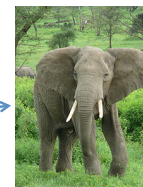
Lexical Similarity: semiotic aspects

Slon (cz)

???

Elefant (dn)

Elefante (pt)



The proposed LS metric

- In our technique we assign more value to the diagrammatic role of consonants than to other matchings and emphasize the role of initials.

The equation for intersymbolic similarity is:

$$IS = \alpha(\gamma_1 I + \gamma_2 C + \gamma_3 V) + \beta P \quad (1)$$

Where: IS: intersymbolic similarity (maximum =1, minimum = 0)

I: initials

C: consonants

V: vowels

P: phonemes (can be decomposed as the orthographical part: $\gamma_4 I + \gamma_5 C + \gamma_6 V$)

α : weight of the orthographical similarity (adjusted according to the context)

β : weight of the phonetic similarity (adjusted according to the context)

γ_n : weights of factors of similarity (e.g. $\gamma_1=0.4$; $\gamma_2=0.4$; $\gamma_3=0.2$)

$\alpha + \beta = 1$ and $\gamma_1 + \gamma_2 + \gamma_3 = 1$ and $\gamma_4 + \gamma_5 + \gamma_6 = 1$

The proposed LS metric

- Weights are adjusted so that the maximum similarity is 1 (totally matching words) and the minimum is 0 (totally different words).
- It may be necessary to normalize consonants and clusters to a same notation: for instance, “š”, “š̂” and “sch” to “sh”.
- The comparisons of the consonant or vowel sequences consider letter groupings such as “cntrl” or “oo”.

The proposed LS metric

Example: The intersymbolic similarities of the Italian word “tempo” respectively to speakers of Portuguese, Spanish, English, German and Finnish are:

L1 (tempo)→L2 (tempo): Initials: t=t; Consonants: tmp=tmp; Vowels: eo=eo
 $IS = 0.6*(0.4*1+0.4*1+0.2*1)+0.4*1 = 1$

L1 (tempo)→L2 (tiempo): Initials: t=t; Consonants: tmp=tmp; Vowels: eo=ieo
 $IS = 0.6*(0.4*1+0.4*1+0.2*0.66)+0.4*0.9 = 0.92$

L1 (tempo)→L2 (time): Initials: t=t; Consonants: tmp=tm; Vowels: eo=ie
 $IS = 0.6*(0.4*1+0.4*0.66+0.2*0)+0.4*0.4 = 0.48$

L1 (tempo)→L2 (Zeit): Initials: t=Z(ts); Consonants: tmp=Zt; Vowels: eo=ei
 $IS = 0.6*(0.4*0.5+0.4*0.16+0.2*0.33)+0.4*0.2 = 0.28$

L1 (tempo)→L2 (aika): Initials: t≠a; Consonants: tmp≠k; Vowels: eo≠aia
 $IS = 0.6*(0.4*0+0.4*0+0.2*0)+0.4*0 = 0$

The proposed LS metric

Original word: “physics” transformations
 to Czech “fyzyka” (sisssss) LD=13
 to Polish “fyzika” (sixsxss) LD=9
 to Afrikaans “fisika” (sisxxss) LD=9
 to Italian “fisica” (sisxxxx) LD=7
 to French “physique” (xxxxxssi) LD=5

The results for intersymbolic similarity are:

$$IS_1 = 0.6*(0.4*0.8 + 0.4*0.65 + 0.2*0.8) + 0.4*0.8 = 0.764$$

$$IS_2 = 0.6*(0.4*0.8 + 0.4*0.65 + 0.2*0.9) + 0.4*0.8 = 0.776$$

$$IS_3 = 0.6*(0.4*0.8 + 0.4*0.72 + 0.2*0.8) + 0.4*0.8 = 0.781$$

$$IS_4 = 0.6*(0.4*0.8 + 0.4*0.80 + 0.2*0.8) + 0.4*0.8 = 0.800$$

$$IS_5 = 0.6*(0.4*1.0 + 0.4*0.90 + 0.2*0.9) + 0.4*0.8 = 0.884$$

The proposed LS metric

- Whereas LD measured distances ranging from 5 to 13, the IS produced similar scores for the five L2 words, arguably because the technique can capture the fact that all words are more or less recognizable based on the original word.
- Conversely, an opposite situation in which two words produce smaller LD, but score worse on IS, would be: “glamour” (en) and “amour” (fr), whose LD=2 is smaller, but whose IS=0.52 indicates less actual similarity.

Conclusions

- We believe that the IS captures the crucial features that make a word more easily recognizable by learners.
- We can assume that there is a threshold below which the recognition will no longer be possible (based on IS).
- A field study is being designed to investigate how this threshold relates to the lexicon of each subject’s L1 and to other known L2s.

Conclusions

- This technique is aimed to offer a practical word-level similarity metric to compare words from different languages so that this measure can be used as an input to initialize the LM or to evaluate word-level errors in the context of CALL applications. It is not aimed to replace other formalisms, neither to create new computational treatments of lexical rules.