Automation and Evaluation of the Keyword Method for Second Language Learning

Gözde Özbal
Trento RISE
Trento, Italy
gozbalde@gmail.com

Daniele Pighin
Google
Zürich, Switzerland
biondo@google.com

Carlo Strapparava
FBK-irst
Trento, Italy
strappa@fbk.eu

Abstract

In this paper, we combine existing NLP techniques with minimal supervision to build memory tips according to the keyword method, a well established mnemonic device for second language learning. We present what we believe to be the first extrinsic evaluation of a creative sentence generator on a vocabulary learning task. The results demonstrate that NLP techniques can effectively support the development of resources for second language learning.

1 Introduction

The keyword method is a mnemonic device (Cohen, 1987; Thompson, 1987) that is especially suitable for vocabulary acquisition in second language learning (Mizumoto and Kansai, 2009; Hummel, 2010; Shen, 2010; Tavakoli and Gerami, 2013). In this method, a *target* word in a foreign language L2 can be learned by a native speaker of another language L1 in two main steps: 1) one or more L1 words, possibly referring to a concrete entity, are chosen based on orthographic or phonetic similarity with the target word; 2) an L1 sentence is constructed in which an association between the translation of the target word and the keyword(s) is established, so that the learner, when seeing or hearing the word, immediately recalls the keyword(s). To illustrate, for teaching the Italian word *cuore* which means *heart* in English, the learner might be asked to imagine *"a lonely* heart *with a hard* core*"*.

The keyword method has already been proven to be a valuable teaching device. However, the preparation of the memorization tips for each new word is an activity that requires considerable time, linguistic competence and creativity. To the best of our knowledge, there is only one study which attempts to automate the mechanism of the keyword method. In (Özbal and Strapparava, 2011), we proposed to automate the keyword method by retrieving sentences from the Web. However, we did not provide any evaluation to demonstrate the effectiveness of our approach in a real life scenario. In addition, we observed that retrieval poses severe limitations in terms of recall and sentence quality, and it might incur copyright violations.

In this paper, we overcome these limitations by introducing a semi-automatic system implementing the keyword method that builds upon the keyword selection mechanism of Özbal and Strapparava (2011) and combines it with a state-of-the-art creative sentence generation framework (Özbal et al., 2013). We set up an experiment to simulate the situation in which a teacher needs to prepare material for a vocabulary teaching resource. According to our scenario, the teacher relies on automatic techniques to generate relatively few, high quality mnemonics in English to teach Italian vocabulary. She only applies a very light supervision in the last step of the process, in which the most suitable among the generated sentences are selected before being presented to the learners. In this stage, the teacher may want to consider factors which are not yet in reach of automatic linguistic processors, such as the evocativeness or the memorability of a sentence. We show that the automatically generated sentences help learners to establish memorable connections which augment their ability to assimilate new vocabulary. To the best of our knowledge, this work is the first documented extrinsic evaluation of a creative sentence generator on a real-world application.

2 Related work

The effectiveness of the keyword method (KM) is a well-established fact (Sarıçoban and Başıbek, 2012). Sommer and Gruneberg (2002) found that using KM to teach French made learning easier and faster than conventional methods. Sagarra and Alba (2006) compared the effectiveness of

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 352–357,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

three learning methods including the semantic mapping, rote memorization (i.e., memorization by pure repetition, with no mnemonic aid) and keyword on beginner learners of a second language. Their results show that using KM leads to better learning of second language vocabulary for beginners. Similar results have been reported by Sarıçoban and Başıbek (2012) and Tavakoli and Gerami (2013). Besides all the experimental results demonstrating the effectiveness of KM, it is worthwhile to mention about the computational efforts to automate the mechanism. In (Özbal and Strapparava, 2011) we proposed an automatic vocabulary teaching system which combines NLP and IR techniques to automatically generate memory tips for vocabulary acquisition. The system exploits orthographic and phonetic similarity metrics to find the best L2 keywords for each target L1 word. Sentences containing the keywords and the translation of the target word are retrieved from the Web, but we did not carry out an evaluation of the quality or the coverage of the retrieved sentences. In Özbal et al. (2013) we proposed an extensible framework for the generation of creative sentences in which users are able to force several words to appear in the sentences. While we had discussed the potentiality of creative sentence generation as a useful teaching device, we had not validated our claim experimentally yet. As a previous attempt at using NLP for education, Manurung et al. (2008) employ a riddle generator to create a language playground for children with complex communication needs.

3 Memory tip generation

Preparing memory tips based on KM includes two main ingredients: one or more keywords which are orthographically or phonetically similar to the L2 word to be learned; and a sentence in which the keywords and the translation of the target L2 word are combined in a meaningful way. In this section, we detail the process that we employed to generate such memory tips semi-automatically.

3.1 Target word selection and keyword generation

We started by compiling a collection of Italian nouns consisting of three syllables from various resources for vocabulary teaching including `http://didattica.org/italiano.htm` and `http://ielanguages.com`, and produced a list of 185 target L2 words. To generate the L1 keywords for each target word, we adopted a similar strategy to Özbal and Strapparava (2011). For each L2 target word t, the keyword selection module generates a list of possible keyword pairs, K. A keyword pair $k \in K$ can either consist of two non-empty strings, i.e., $k = [w_0, w_1]$, or of one non-empty and one empty string, i.e., $w_1 = \epsilon$. Each keyword pair has the property that the concatenation of its elements is either orthographically or phonetically similar to the target word t. Orthographic and phonetic similarity are evaluated by means of the Levenshtein distance (Levenshtein, 1966). For orthographic similarity, the distance is calculated over the characters in the words, while for phonetic similarity it is calculated over the phonetic representations of t and $w_0 + w_1$. We use the CMU pronunciation dictionary[1] to retrieve the phonetic representation of English words. For Italian words, instead, their phonetic representation is obtained from an unpublished phonetic lexicon developed at FBK-irst.

3.2 Keyword filtering and ranking

Unlike in (Özbal and Strapparava, 2011), where we did not enforce any constraints for selecting the keywords, in this case we applied a more sophisticated filtering and ranking strategy. We require at least one keyword in each pair to be a content word; then, we require that at least one keyword has length ≥ 3; finally, we discard pairs containing at least one proper noun. We allowed the keyword generation module to consider all the entries in the CMU dictionary, and rank the keyword pairs based on the following criteria in decreasing order of precedence: 1) Keywords with a smaller orthographic/phonetic distance are preferred; 2) Keywords consisting of a single word are preferred over two words (e.g., for the target word *lavagna*, which means *blackboard*, *lasagna* takes precedence over *love* and *onion*); 3) Keywords that do not contain stop words are preferred (e.g., for the target word *pettine*, which means *comb*, the keyword pair *pet* and *inn* is ranked higher than *pet* and *in*, since *in* is a stop word); 4) Keyword pairs obtained with orthographic similarity are preferred over those obtained with phonetic similarity, as learners might be unfamiliar with the phonetic rules of the target language. For example, for the target word *forbice*, which means *scissors*,

[1] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

Group	Target	Sentence
A1	campagna	a *company* runs the **country**
A1	isola	an *island* of remote **isolated** communities
A1	fabbrica	a fabric worker in a **factory**
A1	bagnino	**lifeguard**s carry *no bag*
A1	inverno	the *inferno* started, **winter** left
A1	cielo	the **sky** has no *ceiling*
A1	marrone	blood and *marrow* in a **brown** water
A1	cuore	the lonely **heart** has hard *core*
A1	coperta	a piece of *copper* in the corner of a **blanket**
A1	locanda	an **inn** oak door with *lock and* key
A2	piazza	a **square** building serves a free *pizza*
A2	calzino	big bloke with **sock** in the *casino*
A2	scatola	a cardboard **box** sat in a *scuttle* of a house
A2	ragazzo	**boy**s *also* have *rag* dolls
A2	angolo	a **corner** kick came at an *angle*
A2	cestino	a *teen* movie uses **basket** to play the *chess*
A2	carbone	the **coal** is the form of *carbon*
A2	cassetto	a blank *cassette* tape is in a **drawer**
A2	farfalla	the **butterflies** are *far* in the *fall*
A2	tovaglia	a damp **cloth** *towel*
B1	duomo	the old **cathedral** has a *dome*
B1	aceto	a **vinegar** sauce contains the *acid*
B1	nuvola	the sophisticated *novel* depicts the **cloud**
B1	chiesa	the Catholic **church** has Swiss *cheese*
B1	bacino	the explosion *in* the *back* broke the *pelvis*
B1	maiale	a **pork** meat comes in the *mail*
B1	minestra	Chinese *ministries* have **soup**
B1	estate	this *estate* is for **summer**
B1	bozzolo	a *buzz* comes wrapped in the **cocoon**
B1	arnese	*harness* a technology to develop a **tool**
B2	asino	an *Asian* elephant is riding a **donkey**
B2	miele	do not make **honey** to walk a *mile*
B2	polmone	crowded *pullman*s stop the **lung**s
B2	fagiolo	a topical *facial* **bean** cream
B2	fiore	a *fire* in a **flower** market
B2	compressa	the clay **tablet** is in the *compress*ed form
B2	cavallo	**horse** running fast in *cavalry*
B2	fiume	the muddy **river** has smoke and *fume*s
B2	pittore	a famous **painter** has precious *picture*s
B2	manico	*manic* people have broken **neck**s

Table 1: Sentences used in the vocabulary acquisition experiment.

the keyword pair *for* and *bid* is preferred to *for* and *beach*.

We selected up to three of the highest ranked keyword pairs for each target word, obtaining 407 keyword combinations for the initial 185 Italian words, which we used as the input for the sentence generator.

3.3 Sentence generation

In this step, our goal was to generate, for each Italian word, sentences containing its L1 translation and the set of orthographically (or phonetically) similar keywords that we previously selected. For each keyword combination, starting from the top-ranked ones, we generated up to 10 sentences by allowing any known part-of-speech for the keywords. The sentences were produced by the state of the art sentence generator of Özbal et al. (2013). The system relies on two corpora of automatic parses as a repository of sentence templates and lexical statistics. As for the former, we combined two resources: a corpus of 16,000 proverbs (Mihalcea and Strapparava, 2006) and a collection of 5,000 image captions[2] collected by Rashtchian et al. (2010). We chose these two collections since they offer a combination of catchy or simple sentences that we expect to be especially suitable for second language learning. As for the second corpus, we used LDC's English GigaWord 5th Edition[3]. Of the 12 feature functions described in (Özbal et al., 2013), we only implemented the following scorers: Variety (to prevent duplicate words from appearing in the sentences); Semantic Cohesion (to enforce the generation of sentence as lexically related to the target words as possible); Alliteration, Rhyme and Plosive (to introduce hooks to echoic memory in the output); Dependency Operator and N-gram (to enforce output grammaticality).

We observed that the sentence generation module was not able to generate a sentence for 24% of the input configurations. For comparison, when we attempted to retrieve sentences from the Web as suggested in Özbal and Strapparava (2011), we could collect an output for less than 10% of the input configurations. Besides, many of the retrieved sentences were exceedingly long and complex to be used in a second language learning experiment.

3.4 Sentence selection

For each L1 keyword pair obtained for each L2 target word, we allowed the system to output up to 10 sentences. We manually assessed the quality of the generated sentences in terms of meaningfulness, evocativeness and grammaticality to select the most appropriate sentences to be used for the task. In addition, for keyword pairs not containing the empty string, we prioritized the sentences in which the keywords were closer to each other. For example, let us assume that we have the keywords *call* and *in* for the target word *collina*. Among the sentences "*The girl received a call in the bathroom*" and "*Call the blond girl in case you need*", the first one is preferred, since the keywords are closer to each other. Furthermore, we gave priority to the sentences that included the keywords

[2] http://vision.cs.uiuc.edu/pascal-sentences/

[3] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2011T07

in the right order. To illustrate, for the same keywords and the target words, we would prefer the sentence "*I called him in the morning yesterday*" over "*You talk a lot in a call*".

Accordingly, for each target word in random order, we sequentially scanned the outputs generated for each keyword pair. As soon as a sentence of adequate quality was found, we added it to our evaluation data and moved on to the next keyword. We continued this process until we selected a sentence for 40 distinct target words, which we set as the target size of the experiment. We had to inspect the outputs generated for 48 target words before we were able to select 40 good examples, meaning that for 17% of the target words the sentence generator could not produce a sentence of acceptable quality.

4 Experiment setup

For our experiment, we drew inspiration from Sagarra and Alba (2006). We compared the retention error rate of learners who tried to memorize new words with or without the aid of the automatically generated sentences. Through academic channels, we recruited 20 native English speakers with no prior knowledge of Italian.[4]

After obtaining the sentences as explained in Section 3, we shuffled and then divided the whole set including 40 target words together with their translation, the generated keywords and sentences into 2 batches (A, B) and further divided each batch into 2 groups consisting of 10 elements (A1, A2, B1 and B2). The set of sentences assigned to each group is listed in Table 1: Column "*Target*" reports the Italian target word being taught; Column "*Sentence*" shows the automatically generated sentence, where the translation of the target word is shown in bold and the keyword(s) in italic. For the experiments, we randomly assigned each subject to one of the batches (A or B). Then, each subject was asked to memorize all the word pairs in a batch, but they would see the memory tips only for one of the two groups, which was again randomly assigned. This approach resulted in 4 different memorization exercises, namely 1) A1 with tips and A2 without, 2) A2 with tips and A1 without, 3) B1 with tips and B2 without, 4) B2 with tips and B1 without.

[4]We preferred to select the experiment subjects in person as opposed to crowdsourcing the evaluation to be able to verify the proficiency of the subjects in the two languages and to ensure the reliability of the outcome of the evaluation.

	Error rate (%)		Reduction	
Group	Rote	KW	Δ_e	$\%_e$
A1	4.08	3.39	0.69	16.95
A2	12.07	10.42	1.65	13.69
B1	12.77	10.00	2.77	21.67
B2	22.50	12.50	10.00	44.44
Macro-average	12.85	9.08	3.78	29.39
Micro-average	11.27	8.25	3.02	26.76

Table 2: Per-group and overall retention error rate when using rote or keyword-aided (KW) memorization.

When memorizing the translations without the aid of memory tips, the subjects were instructed to focus only on the Italian word and its English translation and to repeat them over and over in their mind. Conversely, when relying on the automatic memory tips the subjects were shown the word, its translation and the generated sentence including the keywords. In this case, the subjects were instructed to read the sentence over and over trying to visualize it.

After going through each set of slides, we distracted the subjects with a short video in order to reset their short term memory. After that, their retention was tested. For each Italian word in the exercise, they were asked to select the English translation among 5 alternatives, including the correct translation and 4 other words randomly selected from the same group. In this way, the subjects would always have to choose among the words that they encountered during the exercise.[5] We also added an extra option "*I already knew this word*" that the subjects were instructed to select in case they already knew the Italian word prior to taking part in the experiment.

5 Experiment results

Table 2 summarizes the outcome of the experiment. The contribution of the automatically generated sentences to the learning task is assessed in terms of error rate-reduction, which we measure both within each group (rows 1-4) and on the whole evaluation set (rows 5-6). Due to the presence of the "*I already knew this word*" option in the learning-assessment questionnaire, the number of the actual answers provided by each subject can be slightly different, hence the difference between macro- and micro-average.

[5]Otherwise, they could easily filter out the wrong answers just because they were not exposed to them recently.

The error rate for each memorization technique t (where $t = \mathrm{R}$ for "Rote memorization" and $t = \mathrm{K}$ for "keyword-aided memorization") is calculated as: $e_t = \frac{i_t}{c_t + i_t}$, where c_t and i_t are the number of correct and incorrect answers provided by the subjects, respectively. The absolute error rate reduction Δe is calculated as the absolute difference in error rate between rote and keyword-aided memorization, i.e.: $\Delta_e = e_\mathrm{R} - e_\mathrm{K}$. Finally, the relative error rate reduction $\%_e$ is calculated as the the ratio between the absolute error rate reduction Δe and the error rate of rote memorization e_R, i.e.,: $\%_e = \frac{\Delta_e}{e_\mathrm{R}} = \frac{e_\mathrm{R} - e_\mathrm{K}}{e_\mathrm{R}}$.

The overall results (rows 5 and 6 in Table 2) show that vocabulary learning noticeably improves when supported by the generated sentences, with error rates dropping by almost 30% in terms of macro-average (almost 27% for micro-average). The breakdown of the error rate across the 4 groups shows a clear pattern. The results clearly indicate that one group (A1) by chance contained easier words to memorize as shown by the low error rate (between 3% and 4%) obtained with both methods. Similarly, groups A2 and B1 are of average difficulty, whereas group B2 appears to be the most difficult, with an error rate higher than 22% when using only rote memorization. Interestingly, there is a strong correlation (Pearson's $r = 0.85$) between the difficulty of the words in each group (measured as the error rate on rote memorization) and the positive contribution of the generated sentences to the learning process. In fact, we can see how the relative error rate reduction $\%_e$ increases from ∼17% (group A1) to almost 45% (group B2). Based on the results obtained by Sagarra and Alba (2006), who showed that the keyword method results in better long-term word retention than rote memorization, we would expect the error rate reduction to be even higher in a delayed post-test. All in all, these findings clearly support the claim that a state-of-the-art sentence generator can be successfully employed to support keyword-based second language learning. After completing their exercise, the subjects were asked to provide feedback about their experience as learners. We set up a 4-items Likert scale (Likert, 1932) where each item consisted of a statement and a 5-point scale of values ranging from (1) [I strongly disagree] to (5) [I strongly agree]. The distribution of the answers to the questions is shown in Table 3. 60% of the subjects acknowledged that the memory tips helped them in the memorization process; 45% found that the sentences were overall correct; 65% confirmed that the sentences were catchy and easy to remember; and 50% found the sentences to be overall witty although the sentence generator does not include a mechanism to generate humor. Finally, it is worth mentioning that none of the subjects noticed that the sentences were machine generated, which we regard as a very positive assessment of the quality of the sentence generation framework. From their comments, it emerges that the subjects actually believed that they were just comparing two memorization techniques.

	Rating (%)				
Question	1	2	3	4	5
Sentences helped	5	20	15	35	25
Sentences are grammatical	-	25	30	35	10
Sentences are catchy	-	25	10	50	15
Sentences are witty	-	25	25	50	-

Table 3: Evaluation of the generated sentences on a 5-point Likert scale.

6 Conclusion and Future Work

In this paper, we have presented a semi-automatic system for the automation of the keyword method and used it to teach 40 Italian words to 20 English native speakers. We let the system select appropriate keywords and generate sentences automatically. For each Italian word, we selected the most suitable among the 10 highest ranked suggestions and used it for the evaluation. The significant reduction in retention error rate (between 17% and 45% on different word groups) for the words learned with the aid of the automatically generated sentences shows that they are a viable low-effort alternative to human-constructed examples for vocabulary teaching.

As future work, it would be interesting to involve learners in an interactive evaluation to understand the extent to which learners can benefit from *ad-hoc* personalization. Furthermore, it should be possible to use frameworks similar to the one that we presented to automate other teaching devices based on sentences conforming to specific requirements (Dehn, 2011), such as *verbal chaining* and *acrostic*.

Acknowledgements

This work was partially supported by the PerTe project (Trento RISE).

References

Andrew D. Cohen. 1987. The use of verbal and imagery mnemonics in second-language vocabulary learning. *Studies in Second Language Acquisition*, 9:43–61, 2.

M.J. Dehn. 2011. *Working Memory and Academic Learning: Assessment and Intervention*. Wiley.

K. M. Hummel. 2010. Translation and short-term L2 vocabulary retention: Hindrance or help? *Language Teaching Research*, 14(1):61–74.

V. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707—710.

R. Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55.

Ruli Manurung, Graeme Ritchie, Helen Pain, Annalu Waller, Dave O'Mara, and Rolf Black. 2008. The Construction of a Pun Generator for Language Skill Development. *Appl. Artif. Intell.*, 22(9):841–869, October.

R. Mihalcea and C. Strapparava. 2006. Learning to laugh (automatically): Computational models for humor recognition. *Journal of Computational Intelligence*, 22(2):126–142, May.

A. Mizumoto and O. T. Kansai. 2009. Examining the effectiveness of explicit instruction of vocabulary learning strategies with Japanese EFL university students. *Language Teaching Research 13*, 4.

Gözde Özbal and Carlo Strapparava. 2011. MEANS: Moving Effective Assonances for Novice Students. In *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI 2011)*, pages 449–450, New York, NY, USA. ACM.

Gözde Özbal, Daniele Pighin, and Carlo Strapparava. 2013. BRAINSUP: Brainstorming Support for Creative Sentence Generation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1446–1455, Sofia, Bulgaria, August. Association for Computational Linguistics.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 139–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

N. Sagarra and M. Alba. 2006. The key is in the keyword: L2 vocabulary learning methods with beginning learners of spanish. *The Modern Language Journal*, 90(2):228–243.

A. Sarıçoban and N. Başıbek. 2012. Mnemonics technique versus context method in teaching vocabulary at upper-intermediate level. *Journal of Education and Science*, 37(164):251–266.

Helen H. Shen. 2010. Imagery and verbal coding approaches in Chinese vocabulary instruction. *Language Teaching Research*, 14(4):485–499.

Steffen Sommer and Michael Gruneberg. 2002. The use of linkword language computer courses in a classroom situation: a case study at rugby school. *'Language Learning Journal*, 26(1):48–53.

M. Tavakoli and E. Gerami. 2013. The effect of keyword and pictorial methods on EFL learners' vocabulary learning and retention. *PORTA LINGUARUM*, 19:299–316.

G. Thompson. 1987. Using bilingual dictionaries. *ELT Journal*, 41(4):282–286. cited By (since 1996)6.