Computational Analysis of French Reborrowing Process for English Loanwords

Zhubo Deng[1], Weijia Shi[1], Pei Zhou[2], Muhao Chen[3], Kai-Wei Chang[1]
[1]Department of Computer Science, University of California Los Angeles
[2]Department of Computer Science, University of Southern California
[3]Department of Computer and Information Science, University of Pennsylvania
dzb1998@g.ucla.edu; {swj0419, kwchang}@cs.ucla.edu;
peiz@usc.edu; muhao@seas.upenn.edu

***Abstract*— Analyzing semantic change of loanwords over time between different languages has been a longstanding sociolinguistic problem. However, few efforts have been put to computational approaches to analyzing the semantics of loanwords. In this paper, we present a new computational method for detecting and tracking the semantic change of loanwords between two languages, specifically for the reborrowing process of loanwords. We trained our model on pre-trained historical bilingual English-French word embeddings aligned with MUSE and proposed two quantitative measures of detecting loanwords reborrowing. The first measure analyzes cosine similarity of one word from two languages in the same year, and it is sensitive to cultural shifts and lexicon reborrowing. The second measure analyzes Pearson correlation from the tendency of the cosine similarity and thus predicts the pattern of these reborrowing loanwords. We show that our model can detect reborrowing loanwords that have been discovered in literature.**

I. Introduction

Loanwords are lexical items adopted from another language and integrated (nativized) in the recipient language [1]. English and French have a lot of lexemes in common. Approximately 45% of the English vocabulary comes from French words, while over fifty thousand English words have their origin in French [2]. Since loanwords are cross-cultural and evolving, it is necessary to analyze loanword changes between a "donor" language and a "recipient" language and capture culture shifts across time [1]. Loanwords detection and analysis is a critical and longstanding linguistic problem. Although Moran [3] has published a toolkit for historical linguistics, they remarked that "Automatic approaches for borrowing detection are still in their infancy in historical linguistics."

In this paper, we focus on one specific type of the loanword change, i.e. reborrowing. Reborrowing is the process where a word travels from one language to another and then back to the originating language in a different form or with a different meaning [4]. Even though English consists of a lot of loanwords from French, their meanings are often different from their French origins. As cultural shifting happens, those loanwords may refer to the same semantic again. Fig. 1 shows the reborrowing process of the word of *film* between English and French, which referred to the same semantic around 1940. Other examples include *interview*, *record*, etc. Analyzing the loanwords semantic change benefits researches on linguistics and cultural evolution. Detecting and tracking culture changes is essential for studies of humanity.

With the help of word embeddings [5], semantic change analysis attracted significant attention. A wide range of studies consists of computational analysis on linguistic variation and linguistic change [1], [6], [7], [8], [9], [10], [11]. Linguistic variation focuses on studies with different domains such as geographic variation of language [6] and language variation across research fields [7], whereas linguistic change focuses on studies with chronological analysis of linguistic shifts. Those studies focus on either different domains but under the same time period, or diachronic analysis within one region. However, these methods can barely apply on analysis of lexicon reborrowing, thus there is not much work on lexicon reborrowing analysis. Analyzing lexicon reborrowing is challenging because it non-trivially requires capturing the semantic changes both along time and across languages. While some studies [1], [8] analyzed cross-lingual loanword variation in the current literature, analysis of loanword change along time has not been tackled yet.

In this paper, we propose a method to analyze loanword change, specifically reborrowing process, between English and French along time. This non-trivially includes two steps of the approach: we first put two domains of language into one embedding space, and after that, we create embedding spaces corresponding to consecutive time periods. We use Multilingual Unsupervised or Supervised word Embeddings (MUSE) model [12] to align embeddings of two languages. We use cosine similarity comparison to capture and track loanword semantic changes. We also propose a method that could detect the reborrowing process of loanwords and also characteristics of that reborrowing period.

II. Modeling

We first define multilingual word embeddings. Let D be a set of languages, and for a language $d_i \in D$, we use V_{d_i} to denote its vocabulary. We use $V = \cup_{d_i \in D} V_{d_i}$ as the global vocabulary, where $V_{d_i} = \{w_{d_i \in D}\}$, and $\mathbf{w}$ denotes word embedding vectors.

2375-9259/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDMW.2019.00175

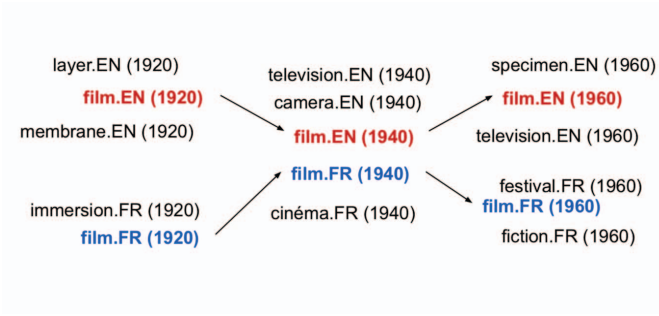

Fig. 1: This figure illustrates the reborrowing process of *film*, an English word (red) of French origin (blue). The words in black are the nearest neighbours of the embeddings of the word *film*. *Flim* referred to *membrane* or layer around 1920 in English, while in French it referred to *immersion*. However, as the cinematic meaning of the word arises in English around 1940, *film* in French borrowed that meaning. Thus, *film* referred to *television* or *camera* both in English and French later around 1940, which indicates the lexicon reborrowing happens. Twenty years later in 1960, the semantics shifted again in both languages.

A. Multilingual Word Embeddings Model

We follow MUSE [3] to align two monolingual word embedding spaces. Since our embeddings are time-variated, for a specific time point t, we align the pretrained English and French word embeddings into one space. Given an English-French dictionary, denoted as V_{d_e} and V_{d_f}, we use the supervised way to learn a mapping from the source space to the target space by using iterative Procrustes alignment and singular value decomposition (SVD):

$$W_t{}^* = \underset{W_t \in O_d(\mathbb{R})}{\arg\min} ||W_t X_t - Y_t||_F = U_t V_t{}^T,$$

with

$$U_t \sum V_t{}^T = SVD(Y_t X_t{}^T),$$

where d is the embedding dimension, $O_d(R)$ is the space of $d \times d$ matrices, and X_t and Y_t are two aligned matrices which contain word embeddings of parallel data from the bilingual dictionary on that specific time point. We get $W^* = \{W_1{}^*, W_2{}^*, ..., W_t{}^*\}$ which are corresponding to t different time spans. By differentiating these time spans, we are able to analyze the semantic change across time.

B. Quantification Approaches

After aligning two languages, we use the following methods to quantify the linguistic change for loanwords in both of the languages.

1) Cosine Similarity:

We consider the cosine similarity between words of two languages in one target space now,

$$\cos(w_i{}^t, w_j{}^t) = \frac{w_i{}^t \cdot w_j{}^t}{||w_i{}^t|| ||w_j{}^t||},$$

where (w_i, w_j) is a word pair given in English-French dictionary, and we use t to denote the current year. As for a pair of reborrowing loanwords, we expect to see a cosine similarity change over the time period. A sharp increase in cosine similarity indicates the reborrowing process to happen during that time period. After that critical time, the semantics may shift due to cultural influences that happened in both languages. Note that here we only consider the cosine similarity in that specific year. Following the definition of reborrowing process, each pair of reborrowing loanwords may undergo a similar pattern, and thus, we need to quantify this pattern.

2) Pearson Correlation:

We quantify the pattern of reborrowing process and use Pearson correlation to detect the reborrowing process. From each sampled reborrowing loanwords pair, we calculate consecutively ten cosine similarity values (ten years as a round, one hundred years in total.) $W_t = \{\cos(w_i{}^1, w_j{}^1), \cos(w_i{}^2, w_j{}^2), ..., \cos(w_i{}^t, w_j{}^t)\}$, where $t_{max} = 10$. Thus, we can track the tendency of semantic change. In each year, we calculate the cosine similarity of each word pair: $S_t = \{\cos(w_{1i}{}^t, w_{1j}{}^t), \cos(w_{2i}{}^t, w_{2j}{}^t), ..., \cos(w_{si}{}^t, w_{sj}{}^t)\}$ We use T, $0 < T < 10$ to denote the duration of the reborrowing process in ten decades. We consider $S = \{S_t, S_{t+1}, ..., S_{t+T-1}\}$ as a set under duration T, and t denotes the reborrowing start time.

In order to calculate the average tendency, for every $S_i \in S$, we take average of elements in S_i, denoted as m_i. Now we have the average tendency $y = \{m_t, m_{t+1}, ..., m_{t+T-1}\}$, and we consider y as a baseline of the reborrowing process. To evaluate roborrowing process of a loanwords pair, we first calculate their cosine similarity W_t, and then we use a sliding window method. We define x_i as the sliding window where i represents the year when the reborrowing process starts, and $10-T+1$ is the window size. For example, if we start with reborrowing period 1900-1950, we slide our window by one unit and we now get to 1910-1960, etc. We keep doing this until reaching the end. To be specific, we calculate $x_k = \{\cos(w_i{}^k, w_j{}^k), \cos(w_i{}^{k+1}, w_j{}^{k+1}), ..., \cos(w_i{}^{k+T-1}, w_j{}^{k+T-1})\}$, where $x_k \in X$, $X = \{x_1, x_2, ..., x_{10-T+1}\}$. Then for every $x \in X$ we can get a Pearson correlation value calculated as follow:

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}.$$

Therefore, we get a sequence of Pearson correlation $R = \{r_1, r_2, ..., r_{10-T+1}\}$. The maximum $r_i \in R$ indicates the period of reborrowing process of that specific loanwords pair. Thus, the reborrowing process can be captured for any random loanwords pair.

III. Evaluation

In this section, we evaluate our methods for analyzing loanwords reborrowing patterns and reborrowing process detection based on a large collection of historical data.

We use the pretrained Historical Word2Vec (SGNS) Embeddings from Stanford HistWords [11], from which we select year from 1900 to 1990, to align our bilingual word embedding space by MUSE. We collect the Reborrowing corpus from The Vocabulary of Modern French [13], and

Term	1920		1950		1980	
	EN	FR	EN	FR	EN	FR
film	layer surface transparent glaze liquid	immersion dante montage hall personnage	camera screen lens pictures surface	cinéma projection théâtre photographie musique	television screen movie emulsion tape	télévision projection festival fiction spectacle
interview	conversation visit announcement asking appointment	times agence journal correspondant radio	questionnaire conversation request interviewer letter	journaliste times match conversation agence	questionnaire survey reporters transcript asked	match reporter express observateur entrevue
record	transcript entries register impressive copy	research proceedings meeting law secretary	document transcript recorder register gramophone	office enterprise verbal document policy	file database document track compile	memory records reporter authority policy

TABLE I: Nearest neighbourhoods of some loanword pairs.

get a list of 200 loanwords of French origin which had the reborrowing process between 1900 and 1990.

A. Dataset and Model Configuration

We align pretrained Historical Word2Vec Embeddings bilingual space by MUSE. HistWords was constructed through Google N-grams, and they released embeddings for each decade, from 1800 to 1990. However, we choose ten decades from 1900 to 1990 where enough corpora have been provided to obtain the embeddings of large vocabularies. For each decade, we learn the mapping by aligning English embedding space and French embedding space. Now we have ten aligned bilingual spaces from ten decades.

B. Change of Loanwords

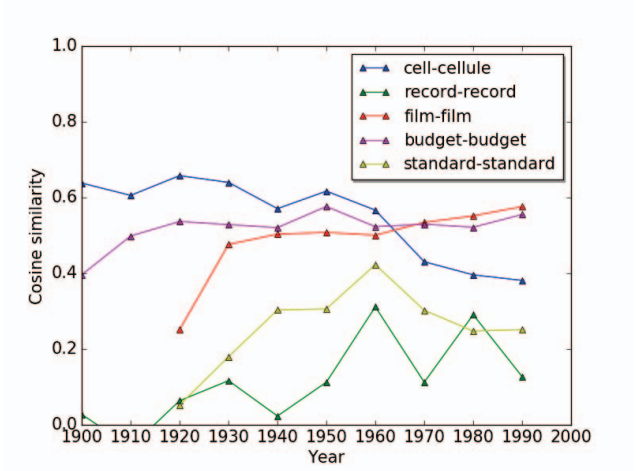

Fig. 2: This figure shows the semantic change of ten loanword pairs from 1900 to 1990.

To evaluate the semantic change of a loanword pair (w_i, w_j), we calculate the cosine similarity of that English-French pair for consecutive ten decades. Specifically, from our aligned ten bilingual mappings, we calculate the cosine similarity on each decade, and get ten cosine similarity values $W_{10} = \{\cos(w_i^1, w_j^1), \cos(w_i^2, w_j^2), ..., \cos(w_i^{10}, w_j^{10})\}$. Fig. 2 shows the ten loanword pairs with their cosine similarity values. Table I shows the nearest neighbourhood changes of three terms, *film*, *interview*, and *record*, which were

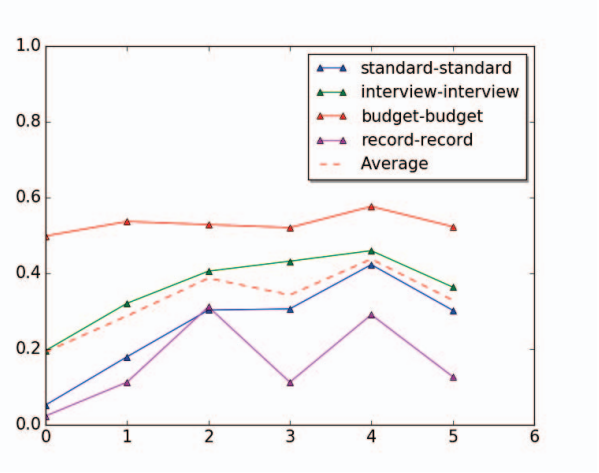

Fig. 3: Cosine similarity value tendency of *standard* (1920-1970), *interview* (1940-1990), *budget* (1910-1960), and *record* (1940-1990). The dash line is the average. (1940-1990)

undergoing the reborrowing process around 1950. The model detects that the meanings of loanword pair *film* in English and *film* in French are getting closer after 1920, as the graph shows that the cosine similarity is getting larger after 1920. Also, as shown in Table I, the cinematic meaning of *film* arises in English around 1950, and *film* in French borrows that meaning. Thus, they have a larger cosine similarity than before. The other example is *record* in English and *record* in French. The documentary meaning of *record* arises in English and is borrowed by French after 1950. If the cosine similarity is decreasing, such as *cell* and *cellule* after 1960, the semantic meaning is moving further from each other during that period of time.

C. Loanword Reborrowing Pattern

In this section, we quantitatively identify the loanword reborrowing pattern. From our reborrowing corpus, we pick four loanwords which had reborrowing process between 1900 and 1990: *standard*, *interview*, *budget*, and *record*. For each word, we manually select the reborrowing time period, based on the literature [13]. Specifically, we select *standard* (1920-1970), *interview* (1940-1990), *budget* (1910-1960), and *record* (1940-1990). Since each reborrowing period has different starting time but similar duration of reborrowing period, our $S = \{S_t, S_{t+1}, ..., S_{t+T-1}\}$ has been adjusted to the same onset t. After that, we take the average of these cosine similarity values, and we get $y = \{m_t, m_{t+1}, ..., m_{t+T-1}\}$. We assume this to be the the baseline of reborrowing process pattern. Fig. 3 shows these reborrowing loanwords with specific time periods, and the average (dash line). The pattern we discovered shows that if a loanword pair has the reborrowing process, then there will be a sharp increase in cosine similarity, and the semantics will shift after a few decades.

Loanwords	$\rho_{Pearson}$	Reborrowing Period	Rank
intellectuals	0.9884	1930-1980	8
genre	0.9795	1920-1970	12
standard	0.9766	1920-1970	26
odour	0.9755	1940-1990	28
interview	0.9750	1940-1990	32
direct	0.9746	1940-1990	35
themes	0.9730	1940-1990	42
budget	0.9676	1910-1960	51
cell	0.9549	1930-1980	64
design	0.9531	1900-1950	73
record	0.9489	1940-1990	77
sports	0.9386	1920-1970	85
management	0.9281	1940-1990	92
film	0.9240	1920-1970	103
humour	0.9213	1930-1980	108

TABLE II: Pearson correlation ranking of loanwords with Reborrowing Process.

D. Reborrowing Process Detection

To verify our proposed reborrowing pattern, we calculate the Pearson correlation with other loanwords. We collect 1000 loanwords of French origin from List of English words of French origin [14], and we try to detect loanwords with reborrowing process form these 1000 words. We use a sliding window method to position the time period of reborrowing process. For each pair of loanword, we calculate the Pearson correlation $10-T+1=5$ times since our sliding window size is $T=6$. We choose the highest Pearson correlation from $R=\{r_1,r_2,...,r_{10-T+1}\}$ which is corresponding to the reborrowing process, and we rank these values for 1000 loanwords. Table II shows that our detection was correct, and the loanwords with reborrowing process have higher rankings than those loanwords which did not have the reborrowing process.

We then calculate the average ranking for 200 loanwords pairs which had reborrowing process, and we compare it with the average ranking of 200 random selected pairs. We get 285.77 for the average ranking of 200 reborrowing pairs. In contrast, we sample 200 loanword pairs for the pre-selected 1000 loanwords five times and receive the average ranking of 508.84 ± 17.91. Our method shows the average ranking of reborrowing loanwords pairs is higher than the average ranking of random selected loanwords pairs, and thus, reborrowing process can be detected by our method.

IV. Related work

In this section, we discuss related work on loanwords analysis, and linguistic change and variation analysis methods.

A. Loanwords Analysis

Few recent work has conducted computational analysis on loanwords [1], [9]. [1] focused on knowledge of morphological and phonological patterns of borrowing words and generation of loanwords pronunciation. Our work is similar to [9], but instead of directly comparing the neighborhoods of loanword pairs from current literature, we aligned bilingual language spaces for ten different time periods. Thus, we can clearly see how semantic changes of a loanword happens over time.

B. Linguistic Change and Variation Analysis

To analyze linguistic change and variation, neural language models got more attention for different tasks. There are two methods mainly used in recent work: diffusion-based and bias-based. Diffusion-based method categorizes different scenarios such as time periods or scientific fields, and generate different word embedding spaces corresponding to these scenarios. Related work includes [10], [7], [9]. Bias-based method represents all words in the same word embedding space but adds a scenario bias vector to the words which appear in that scenario. Related work includes [6], [7]. Since our task focuses on loanword semantic change across time, our work mainly adopted the diffusion-based method, and then we aligned our bilingual word embedding spaces according to different time periods.

V. Conclusion

In this paper, we proposed a computational method for analyzing loanword change and reborrowing detection. The model was trained on a collection of the historical bilingual corpus from ten decades to obtain bilingual embedding spaces. We evaluated the loanwords change by comparing loanwords pair semantics over time. Our baseline for loanword reborrowing pattern shows that it detects the reborrowing process correctly. We believe that with the help of detecting loanword change, the confusion between cross-lingual literature will be reduced, and translation between two languages will have fewer errors on loanword ambiguity.

References

[1] Y. Tsvetkov and C. Dyer, "Cross-lingual bridges with models of lexical borrowing." *JAIR*, 2016.

[2] C. Watkins, *The American Heritage Dictionary of Indo-European Roots.* Mifflin, Houghton, 2000.

[3] J.-M. List and S. Moran, "An open source toolkit for quantitative historical linguistics." in *ACL*, 2013.

[4] P. Durkin, *The Oxford Guide to Etymology.* OUP Oxford, 2009.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space." in *ICLR*, 2013.

[6] V. Kulkarni, B. Perozzi, S. Skiena *et al.*, "Freshman or fresher? quantifying the geographic variation of language in online social media." in *ICWSM*, 2016.

[7] P. Zhou, M. Chen, K.-W. Chang, and C. Zaniolo, "Quantification and analysis of scientific language variation across research fields." *ICDM Workshop*, 2018.

[8] W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Cultural shift or linguistic drift? comparing two computational measures of semantic change." in *EMNLP*, 2016.

[9] H. Takamura, R. Nagata, and Y. Kawasaki, "Analyzing semantic change in japanese loanwords." in *ACL*, 2017.

[10] Y. Kim, Y.-I. Chiu, K. Hanaki *et al.*, "Temporal analysis of language through neural language models." *LTCSS*, 2014.

[11] W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Diachronic word embeddings reveal statistical laws of semantic change." in *ACL*, 2016.

[12] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data." in *ICLR*, 2018.

[13] H. Wise, *The Vocabulary of Modern French.* Routledge, 1997.

[14] S. Zdravkovic, "List of English words of French origin," https://www.ezglot.com/etymologies.php?l=engl2=frasubmit=Compare, 2008.