Exploring Spatial Differences in Vector Representation of Original and Translated Texts

Sangyoon Lee
Graduate School of Culture Technology, KAIST
Republic of Korea
sl2820@kaist.ac.kr

Abstract

A schema refers to a type of cognitive representation that people use to organize and interpret perceptual information of the world. Typically, the schema of a concept is the composition of semantically similar information and enables people to understand the world more efficiently. In the field of machine learning, a vector representation of input data corresponds to the schema of human beings as the computers perceive the information in the numeric form. There have been only a few works which specifically discuss the meanings and applications of the vector representations of different inputs for computers; instead, the majority of works primarily focus on the application of these vectors. This paper aims to explore the meaning of vector representation of text data, embedded by Word2Vec model. Specifically, by comparing the spatial vector representation of the original text and translated text, the paper demonstrates qualitative differences among the vector representations.

Keywords Artificial Intelligence, Word Embedding, Machine Translation, Cognitive Psychology

1. Introduction

When humans absorb different perceptual stimuli from the environment, they organize and comprehend semantically similar information near each other on their minds. When people imagine an orange, not only the color and shape but also a taste of it appear on people's minds simultaneously because these are semantically similar attributes that people naturally associate with an orange; Such mental representation is called "Schema" in the field of psychology.

The schema enables an efficient information processing of the surrounding environment because people can organize new perceptions based on existing schema quickly without requiring extra attention [3, 12]. Given its pivotal role in human cognition, the understanding of such representation has been one of the vital research areas in the historical course on the field of Psychology.

Similarly, computers assimilate input data in numeric representations. A vector is a conventional representation of such data; for instance, for computer, an image is represented as a vector composition of its RGB color values. Lately, the breakthroughs in hardware acceleration using GPUs and algorithmic solutions have brought its pivotal moments. One of the essential outcomes of such advancement is natural language processing. Notably, Machine Translation can deliver naturalness of the language, and even contextual understanding and Speech Recognition enables near hands-free life for people [2,6]. The core algorithm that enabled such performances is a word embedding model called Word2Vec.

Word2Vec is a word embedding algorithm that takes a text corpus data and produces a vector representation for each word from the corpus [10]. It is exceptionally well known for its preservation of words semantically in the vector space [11]. As shown in Figure 1, semantically similar words in the text corpus are located near each other. While such semantic representation has been vigorously appropriated in the advancement of many natural language processing tasks, there has only been a few research that solely explores the meaning of the vector representation itself.

This paper aims to explore the vector representation of text by comparing spatial differences in the vector representation of the original text and translated texts. The paper presents three different geometric analysis to compare the vector representation of original and translated texts. For the analysis, translation is chosen as variation methods because its quality is exclusively depending on the semantic similarities between the target text and source text.

2. Related Works

Word embedding process for a computer is virtually the same process as text cognition for a humans. This section

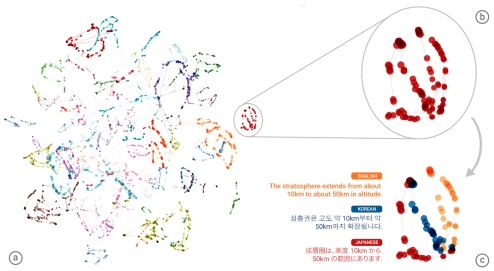

Figure 1. Part (a) from the figure above shows an overall geometry of these translations. Each color represents the meaning. Part (b) zooms in to one of the color groups, and part (c) represents the different languages within the single color groups [8].

discusses Word2Vec model in more detail and one of its main application, translation.

2.1. Word Embedding Model: Word2Vec

There have been many attempts to embed text data into a vector formation. Word2Vec model, developed by Tomas Mikolov from Google, marks one of the most effective word embedding technique available at the moment. The model is a simple two-layered neural network that is trained to reconstruct linguistic contexts of words by producing a different vector for each word, typically in several hundred dimensions [10].

Taken from the distributional hypothesis, Word2Vec demonstrates exceptional performance at capturing syntactic and semantic regularities in the language in terms of vector space [10, 11]. This allows spatial reasoning based on the vector representation between different words. For example, the model captures the semantic relationship of male-female is similar to that of King-Queen and automatically embeds two pairs proportionally in the vector space, as shown in the right image of Figure 2. With such phenomenal preservation of semantics, Word2Vec has brought a significant breakthrough on many of natural language processing tasks. For example, Figure 3 illustrates the similar geometric arrangements of the vectors for numbers and animals in English and Spanish. The authors argued that because most human languages share concepts that are grounded in the real world, there exists a substantial similarity between vector embedding for one language to another [9]. From another practical analysis in 2016, Bolukbasi et al. explored word embedding representation of Word2Vec model and revealed that its semantic relationships are disturbingly biased [1]. According to their study, given word pair of men to women, Word2Vec yielded stereotyped pair of Computer Programmer to Homemaker and Physician to Nurse for its output.

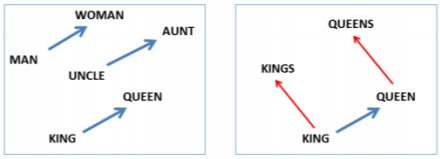

Figure 2. Vector space representation for different pairs of semantically related words. Left figure shows vector geometry for three word pairs illustrating the gender relation. Right figure shows a different projection of the singular/plural relation for two words. In high-dimensional space, multiple relations can be embedded for a single word [11].

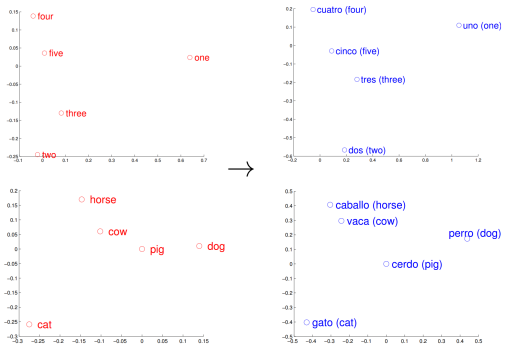

Figure 3. Distributed word vector representations of numbers and animals in English (left) and Spanish (right) [9]

2.2. Machine Translation and Its Quality

The current state of the art machine translations provides reasonable accuracy with incredible efficiency. Widely available free translation tools, such as Google Translate, have given a handy way for anyone with an internet capable device to generate decent translation (or even live-translation) [14].

Several evaluation algorithms are conventionally used for measuring machine translation quality. While the oldest and the most precise evaluation is the use of human judgment, it requires arduous manual works and massive dedication of time. Hence, several automated evaluation algorithms have introduced recently. Among them, BLEU (bilingual evaluation understudy) is the most predominantly used standard. The quality measures are calculated for individual segments of translation by comparing them with high-quality reference translations [7].

3. Model Description

This section describes the overall process of analysis. For this paper, novel Demian, by Hermann Hesse, was selected for exploring the geometry of vector representation [4].

3.1. Data Preparation

In order to explore the vector representation of text data, the study required representations from the source text and target text. Given original German version of novel *Demian* as source text, there were three types of target texts for *Demian*; the official version of English translation [5] and two machine translation versions, Google Translate version of an English translation, and Microsoft Translate version of an English translation.

The official version of the English translation is collected from free online archives. Unfortunately, the machine translation version was not available on the online, and two machine translation versions are manually translated by the author, using Google Translate and Microsoft Translate websites. Given the 5000 characters limitation, the original German version has to be translated in multiple chunks. To minimize the loss of meanings over the translation, the size of each chunk was maintained as much as possible. Because of the difference in each algorithm's counting of backspace and enter keys, a total of 73 chunks were used for Google Translate, and 75 chunks were used for Microsoft Translate.

3.2. Word Embedding: Word2Vec

All the analysis and model implementations are done via Google Colab Notebook, a free Jupyter notebook environment that requires no setup and runs entirely in the cloud as a script. One of the most frequently used open source library, Gensim, is utilized for implementation of Word2Vec model. Each text input goes through the model and comes out as a high dimensional vector in several thousand by 300. The first column represents the number of words in the input text corpus, and it varies among the input text ranging from 5000 words to 8000 words. The second column represents the dimensionality for a word, and the study set it to be 300 dimensions. This means that different values in 300 dimension vector uniquely represent each word.

Typically, a similarity between two words is calculated via cosine distance between two vectors. However, as the mere value of the distance is very obscure to compare, the current study focuses on the geometric representation of word vectors in a two-dimensional plot. To do so, the output vector of several thousand by 300 is reduced into two dimensions. T-Distributed Stochastic Neighboring Embedding (T-SNE) is used for nonlinear dimensionality reduction technique.

3.3. Geometric Relationships of Words

As a vector representation of text input captures the semantic relationships among words geometrically [9], this study compares the spatial representation of different words among German version of *Demian* and translated version of *Demian*. Specifically, the positions of words and directions of one word to another are used to measure the differences.

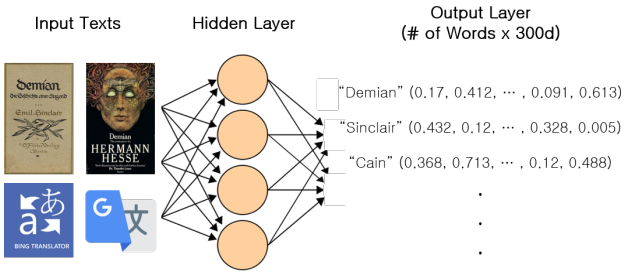

Figure 4. Data Flow for Vector Representation. Text input goes through the hidden layer of Word2Vec model and the model produces vectors of each word from text input.

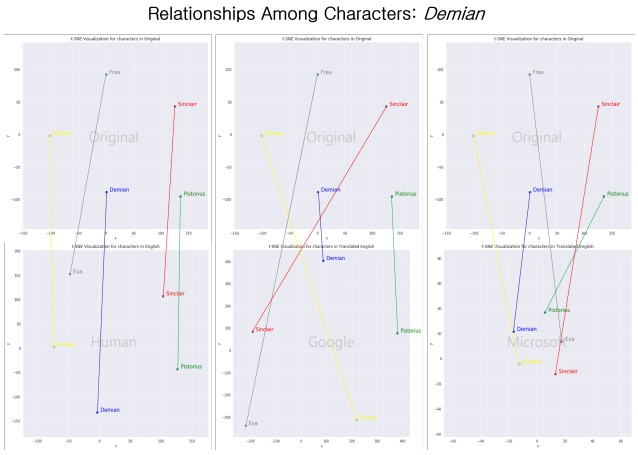

Figure 5. Vector Representations of Main Characters in Novel *Demian*. Original texts are at the top row and translated texts are at the bottom row.

The positions of words indicate how the words are located in the vector space; for example, given words A, B, and C, they may be positioned in the order of A, B, and C or of B, C, and A. The directions of words indicate the directions of words starting from a selected target word. These analyses are based on the assumption that positions of words are semantically meaningful, and the study explores the similarity of the spatial vector representations of different text inputs.

4. Results

One critical assumption for the upcoming analysis is that human translation would preserve semantic relationships among different words much better than the machine translation despite the recent innovations in its capability. This assumption will be useful for qualitative comparison in the current study.

Spatial Relationships Among Characters

First, spatial relationships among the main protagonists were analyzed. Five characters, Demian, Sinclair, Frau Eva,

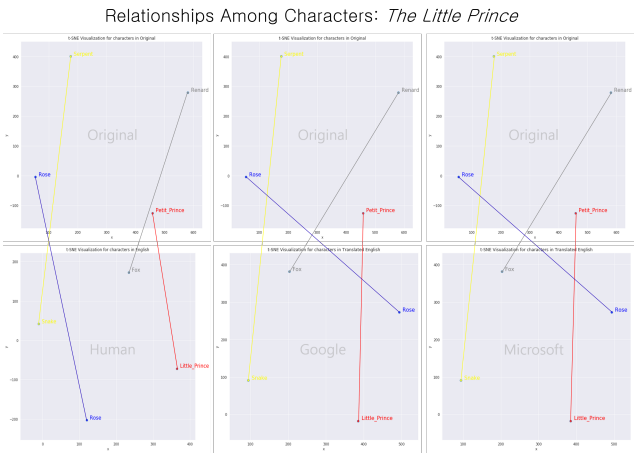

Figure 6. Vector Representations of Main Characters in Novel "The Little Prince." Original texts are at the top row and translated texts are at the bottom row.

Kramer, and Pistorius, were used for vector representation. In Figure 5, the three images on the top row are the same vector representation of five characters from the original German text. The bottom three plots are the vector representation of English translated texts, by Human, Google Translate, and Microsoft Translate from left to right, respectively. The colored lines connect the same characters from the original text representation (top row) to translated text representations (bottom row). The lines are connected to show the positional preservation between the original text and translated text; if the lines are more parallel to each other, the semantic relationships are better preserved.

As visually noticeable from Figure 5, while the line connections between the original text and human translated text are relatively parallel, the line connections between original text and machine translated texts are overlapping one another. This shows that human translation tends to preserve the characters spatially similar to the original texts than the other two machine translations.

To evaluate whether the interpretation of such finding is a general phenomenon, another novel, The Little Prince, has been embedded and analyzed. Surprisingly, a similar pattern appeared from the character analysis for the Little Prince text as well. From Figure 6, although human translation does have colored lines crossing over, it is not as much as the overlaps of colored lines in machine translated texts. Therefore, it can be interpreted that machine translated texts may not represent the relationship among characters as closely as the human translated text does.

Semantic Relationships Among Characters

As described previously, the Word2Vec preserves semantic relationships among the words-pairs, like King-Queen to Prince-Princess, on the vector space geometrically. In the second analysis, semantic relationships among characters are investigated in a similar manner.

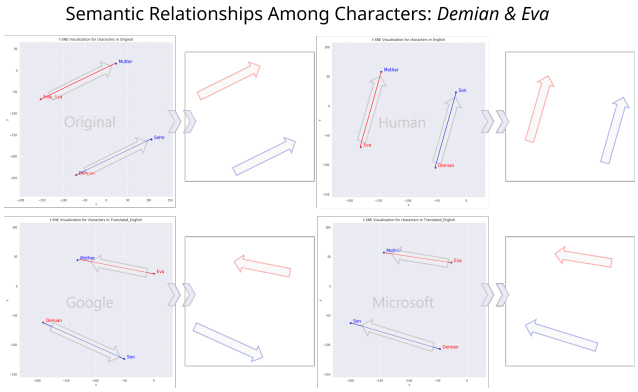

Figure 7. Vector Representations of Semantically Related Word Pairs, Demian and Eva and Son and Mother.

In the novel, one of the most apparent relationships is between Demian and Frau Eva, where Eva is the mother of Demian. Hence, four words, *Demian*, *Eva*, *Son*, and *Mother*, are projected in the vector space for each text input. In Figure 7, the red arrow denotes the connection of words *Eva* and *Mother*, and the blue arrow represents the connection of words *Demian* and *Son*. The direction of arrows indicates the pointing from character to its semantic word, such as *Eva* to *Mother* and *Demian* to *Son*. The top-left plot is the vector representation of the original German text. The human translated text is on the top-right plot, and the bottom two plots are the machine translated texts, by Google Translate and Microsoft Translate from left to right, respectively.

Although it does not maintain the direction of arrows precisely, the human translated text representation resembles the most reasonable and geometrically close to the representation of original text; the arrows of two machine translated texts are pointing to the entirely different directions.

To further assess, another semantic relationship from the novel, *Demian* and *Sinclair* with *Teacher* and *Student*, was analyzed; because, in the novel, Demian acts as a guide and teaches Sinclair about the real world.

Unfortunately, as shown in Figure 8, the positions and directions of arrows from all three translated texts are completely unrelated to those from the original text. One possible conjecture for such discrepancy was that unlike, *Demian* and *Eva*, where the relationship of *Mother* and *Son* is explicitly described in the novel, the relationship between *Demian* and *Sinclair* as *Teacher* and *Student* is more of defined relationship, which is implicitly portrayed.

Character-Keywords Relationships

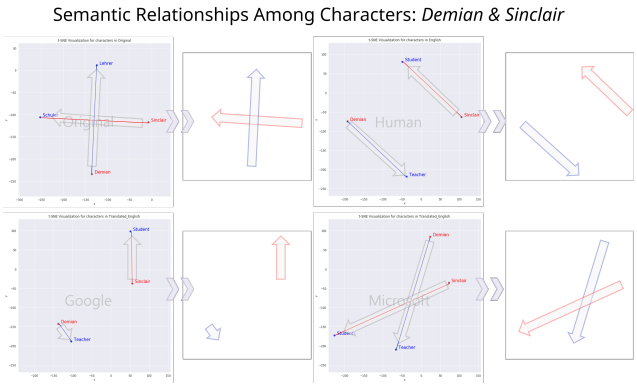

Figure 8. Vector Representations of Semantically Related Word Pairs, Demian and Sinclair and Teacher and Student.

The final analysis explores how keywords related to a specific character are spatially represented in the vector space.

Three keywords for a particular protagonist were manually selected based on the characteristics of the protagonist in the overall storyline. For Sinclair, keywords, *Worry*, *Desire*, and *Fear* were selected. For Demian, keywords, *Leader*, *Ideal*, and *Hope* were chosen. In Figure 9 and 10, the spatial relationship between keywords and characters are shown. Colors are used to separate the different keywords, and the direction of arrows is pointing from the position of the character to the positions of the keywords. The top-left plot is the representation of the original German text. The human translated text is on the top-right plot, and the bottom two plots are the machine translated texts, by Google Translate and Microsoft Translate from left to right, respectively.

For Sinclair and keywords in Figure 9, in the vector representation of original text, *Desire* is in the middle, *Worry* is on the left, and '*Fear* is on the right. However, for all three translated texts, *Worry* is in the middle, *Fear* is on the left, and *Desire* is on the right. Further, the direction of the arrows from translated texts is not congruent to that of the original text neither.

Then, for Demian and keywords in Figure 10, it appears that directions of arrows are all pointing in comparable directions for both the original text and translated texts, yet, the positions of keywords are different. Original text has *Ideal* in the middle, with *Leader* on the left and *Hope* on the right. On the other hand, for all three translated texts, *Ideal* was also in the middle, yet *Hope* is on the left and *Leader* is on the right.

Overall, the keywords-character relationship is not accurately preserved across the translations. As for both Sinclair and Demian cases, the directions of arrows and positions of keywords were hardly consistent between original texts and three different translated texts. However, one thing to mention is that, interestingly, all three translated representations are related in terms of directions of arrows as well as the positions of the keywords in the vector space. One possible, but careful speculation is that it could be due to the handling of synonyms in translation. For instance, word *Angst* could be translated into *Fear*, *Anxiety*, or any other word in the similar meaning. Hence, if a better selection of keywords is possible, this issue might be avoided.

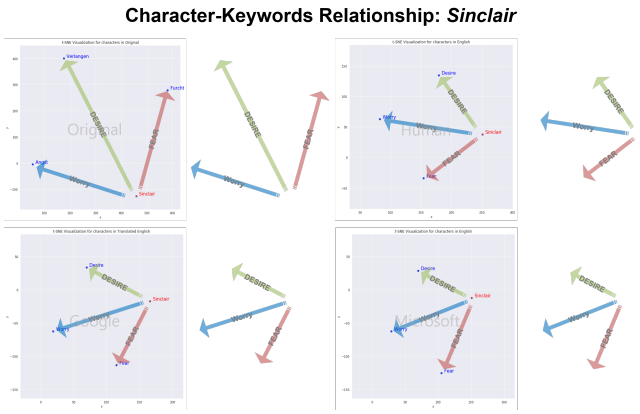

Figure 9. Vector Representations of Sinclair and Semantically Related Words: Worry, Desire, and Fear.

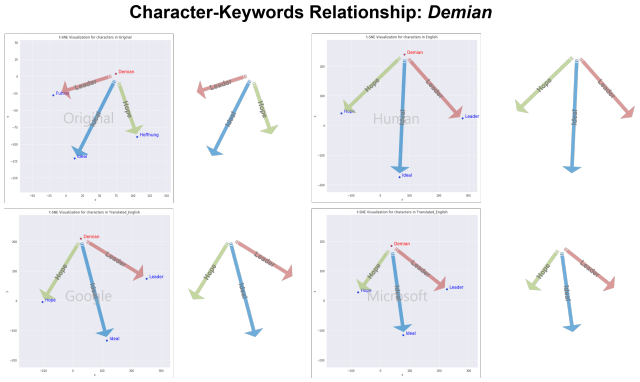

Figure 10. Vector Representations of Demian and Semantically Related Words: Leader, Ideal, and Hope.

5. Discussion

5.1. Summary and Contribution

Typically, the vector representation of text has been primarily used for the pre-processing of texts before applied to other applications like Machine Translation or Speech Recognition. Instead, this study explored the text data in the vector space itself, and there were several interesting discoveries on the vector representation of text data.

As first and second analysis demonstrated, the vector space embeds the names of characters in meaningful posi-

tions when a source language text is translated into a target language text accurately; human translation exceptionally maintained the positions and directions of characters much better than those of machine translation. While some studies showed semantic preservation of common vocabularies such as countries and capitals or present tense verb and past tense verb, analysis on the names of characters is first of its kind.

Furthermore, the study revealed that the similarity in the geometric representation of text does have meaningful information. As shown from the three analysis, the vector representations between semantically similar texts, original text, and human translation, preserve analogous geometries.

5.2. Limitation

The current study raises several critical limitations.

Among them, the most severe limitation is qualitative measurements for geometric relationships in the vector space. Majority of the analysis considers the visual representation of the data. Then, instead of measuring specific values, this study only presents a verbal description of the overall geometry. Although this is a unique investigation, it leaves subjective and controversial interpretations and requires more quantitative measures such as measuring the areas or angles of vector positions in the space.

Moreover, the current study uses the simple version of Word2Vec model. Since its first publication in 2013, there have been many extensions of Word2Vec model. For instance, one model called *GloVe* extends from Word2Vec model and provides better performance in the word analogy tasks, similarity tasks, and named entity recognition tasks [13]. Hence, if the better model is applied, the analysis may produce more accurate results.

Finally, the current study probes only a single novel. In order to generalize the results, more novels, and even other types of text data, such as news article or movie scripts, should be applied.

6. Conclusion

The study begs for further research and development on the idea of using spatial representation in vector space. As mentioned in the discussion, the comparison methods presented in this study are the first of its kind. While there are various limitations, the study provided a reasonable analysis of vector representations of words in the novel. Geometries of the vector representation, such as directions between words and positions of words, do demonstrate semantic preservations.

Hence, if the limitations can be solved, there could be several future works that are promising. For instance, quality of translation be measured by comparing vector geometry of keywords in the source text and target text; for novels, names of the characters can be used, and for news articles, main keywords or hashtags can be used.

As society advances, the integration with A.I. enabled technologies is inevitable in the near future. However, many of the current research solely focuses on the performances of the technologies. Just as the current study tried, there should be more efforts in understanding how machines perform. These research may give a hint to the better interaction with machines in the future.

References

[1] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520, 2016.

[2] L. Deng and D. Yu. Deep learning: Methods and applications. Technical Report MSR-TR-2014-21, May 2014.

[3] P. DiMaggio. Culture and cognition. *Annual Review of Sociology*, 23(1):263–287, 1997.

[4] H. Hesse. *Demian: the story of Emil Sinclair's youth*. Penguin Classics, 2013.

[5] H. Hesse. Demian. Apr 2015.

[6] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, et al. Googles multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[7] N. Madnani. ibleu: Interactively debugging and scoring statistical machine translation systems. In *2011 IEEE Fifth International Conference on Semantic Computing*, pages 213–214. IEEE, 2011.

[8] M. J. Mike Schuster and N. Thorat. Zero-shot translation with googles multilingual neural machine translation system. *Google Blog, available [online] at https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html*, Nov 2016.

[9] T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[11] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013.

[12] S. Nadkarni and V. K. Narayanan. Strategic schemas, strategic flexibility, and firm performance: The moderating role of industry clockspeed. *Strategic management journal*, 28(3):243–270, 2007.

[13] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[14] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.