https://doi.org/10.2352/issn.2168-3204.2021.1.0.23

Data Operations and an Application for Translating Russian Speech to French Text

Steven Simske; Colorado State University; Fort Collins, CO, USA, and Marie Vans; HP Inc; Fort Collins, CO, USA

Abstract

In 2006, the French government discretely asked for an assessment of the highest accuracy means available at the time to translate Russian speech into French text. One of us was working with the Grenoble HP site at the time, and so promptly assessed the possibilities using existing speech-to-text and translation software (Nuance and Speechworks). This article describes the surprisingly circuitous route to maximum accuracy (90.3%), and in so doing provides an unexpected insight into discerning the native language of software designed for speech-to-text and translation applications.

Introduction

From 2003-2008, one of us (Simske) was working with the HP Open Call Business Unit (OCBU), which provided research and development, support, and connectivity services for telephony throughout France. In late 2006, a representative of the French government discretely asked if there was a way for OCBU to investigate the highest accuracy means of translating Russian Speech into French text. The same one of us asked whether it was related to a diplomatic incident in the United Kingdom the month before [1] and was rewarded with being asked to perform the investigation. This is the story of that investigation. Names of other people involved are not provided in case of their preference for anonymity. To our knowledge, the recommendations provided here were not actually implemented by the French government. However, we cannot rule that out. Given the time that has passed since this work was performed, however, we are confident that the required "quiet time" after performing the work has been satisfied.

Methods and Materials

The speech-to-text (recognition) and text-to-text (translation) software used at the time were Nuance and Speechworks engines. These language software providers have changed substantially in the 15 years since this "throwaway" investigation was performed, but the process outlined here could be performed anew with current engines, if desired. These engines were licensed by the HP Open Call Business Unit (OCBU) at the time, and the settings were determined by the Grenoble, France, HP business unit. The software enabled the ability to translate Russian speech into (a) Russian text, (b) English speech, or (c) French speech, as noted in [2]. The inclusion of and intermediate translation into English speech was included as a possible intermediary step based on earlier findings wherein it was observed that English often served as the "central" language for the recognition and translation software. This centrality of English means that there is often a "high-accuracy pipeline" in the middle of a language translation task based on using English as one or both ends of a step in the overall process.

Once English speech is obtained, it can be transformed into English text with very high accuracy. Russian text can separately be translated into English text or French text. The English text can be translated into the final form, French text. French speech can also be directly transformed into French text. Piecing all of these data operations (recognitions and translations) together, we arrive at the generalized graph shown in Figure 1.

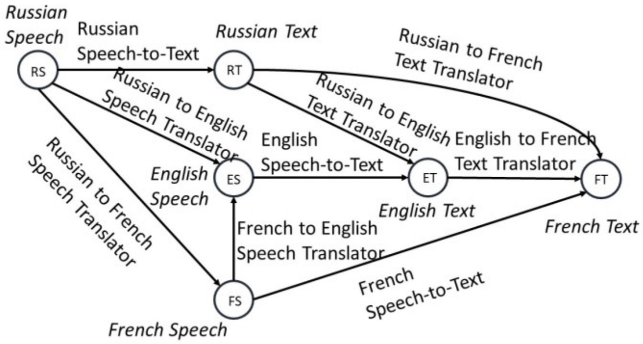
Russian Speech
Russian Speech-to-Text
Russian Text
Russian to French Text Translator
RS
RT
Russian to English Speech Translator
Russian to English Text Translator
English Speech-to-Text
English to French Text Translator
Russian to French Speech Translator
English Speech
ES
ET
FT
English Text
French Text
French to English Speech Translator
FS
French Speech-to-Text
French Speech

Figure 1. Generalized graph (nodes and edges) of a specific text analytics task, in this case translating Russian speech into French text. Because an English speech and English text intermediate exists, there are five relevant paths for translation in the graph shown here (and described in more detail in the text). The uppermost path is from Russian speech to Russian text, and then to French text. The lowermost path is from Russian speech to French speech, and then to French text.

In Figure 1, the generalized graph illustrated consists of nodes that correspond to one of three languages {Russian, English, French} = {R, E, F} and one of two data types {Speech, Text} = {S, T}. Thus, there are six nodes. The edges correspond to the transformations and translations (combined, these are designated "operations") mentioned before, and each of the edges can either change R, E, or F to another language or change S or T to T or S, respectively. A path in Figure 1, therefore, is a set of two or more consecutively traveled edges that convert RS to FT. Five different reasonable paths can be traveled, and they are the basis of the edge directions in Figure 1. These five are:

(1) Russian speech → Russian text → French text, or RS-RT-FT
(2) Russian speech → Russian text → English text → French text, or RS-RT-ET-FT
(3) Russian speech → English speech → English text → French text, or RS-ES-ET-FT
(4) Russian speech → French speech → French text, or RS-FS-FT
(5) Russian speech → French speech → English speech → English text → French text, or RS-FS-ES-ET-FT

In order to assess the highest accuracy pathway for RS→FT, each of the nine edges (RS-RT, RS-ES, RS-FS, RT-FT, RT-ET, FS-ES, ES-ET, ET-FT, and FS-FT) was evaluated on a small data set of sentences (100). Although insufficient for a highly significant statistical assessment, this "throwaway" experiment was sufficient for the evaluation of different approaches to the RS→FT translation and for determining if the software was based on English.

Accuracy was assessed by judgement of three bi-lingual speakers of at least the two languages involved in the particular edge. The process for evaluation using three experts on a linguistic tasks such as this is described elsewhere [3][4].

Results

After determining the accuracies (as probabilities of the correct result being obtained), these are assigned to each edge in the graph. Because of the small number of documents (and a grand total of only 47 inaccurate steps, e.g. 12 in edge RS-ES), assessment of edge to edge correlation was limited, but the 47 errors occurred on 25 documents (range 1-4 errors on the 100 documents for the nine edges), indicating general independence of each edge from the other.

In Figure 2, these accuracy probabilities are placed on each edge, as follows:

(1) Russian speech transformed into Russian text, accuracy = 0.95
(2) Russian speech translated into English speech, accuracy = 0.88
(3) Russian speech translated into French speech, accuracy = 0.94
(4) Russian text translated into French text, accuracy = 0.94
(5) Russian text translated into English text, accuracy = 0.93
(6) French speech transformed into French text, accuracy = 0.93
(7) French speech translated into English speech, accuracy = 0.98
(8) English speech transformed into English text, accuracy = 0.99
(9) English text translated into French text, accuracy = 0.99

The number of errors was therefore given by:

(1) Russian speech transformed into Russian text, 5 errors
(2) Russian speech translated into English speech, 12 errors
(3) Russian speech translated into French speech, 6 errors
(4) Russian text translated into French text, 6 errors
(5) Russian text translated into English text, 7 errors
(6) French speech transformed into French text, 7 errors
(7) French speech translated into English speech, 2 errors
(8) English speech transformed into English text, 1 error
(9) English text translated into French text, 1 error

From these accuracy values, we can see that certain operations in the overall system have extremely high accuracy; namely, any that transform or translate from English. These high accuracies imply that English was likely the language of centrality for the system. Secondly, the accuracy of 0.93 for transformation (6), French speech to French text, shows that French was not likely a language of proficiency for the folks who built all of the different translation/transformation algorithms. Thirdly, the accuracies of translating English speech into French speech, French text into English text, English speech into Russian speech, English text into Russian text, and French speech into Russian speech are not given. We therefore do not know how close these are to their inverse operations, which have accuracies of 0.98, 0.99. 0.88, 0.93, and 0.94, respectively. If they are significantly different from the accuracies of these inverse operations, then these are asymmetric operations, which could be used to give further insight into the linguistic origins of the software.

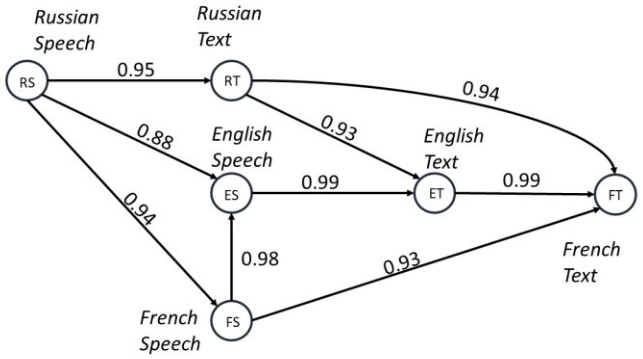

Figure 2. *Generalized graph of Figure1 with the accuracy (probabilities) indicated. If each edge in a sequence of edges traversing from RS to FT is independent of each other, then the accuracy (probability) of the path is simply all of the edge probabilities multiplied together. This assumption drives the values in Table 1.*

Having added these accuracies (as probability values) to the edges of Figure 2, the expected accuracies of each of the five paths from RS to FT can be computed. If the individual operation accuracies are independent of each other, then the path accuracy is simply each of the edge accuracies in the path multiplied together. These are shown in Table 1.

Path	Edge Accuracies	Total Accuracy	Rank
RS-RT-FT	(0.95)(0.94)	0.893	2
RS-RT-ET-FT	(0.95)(0.93)(0.99)	0.875	3
RS-ES-ET-FT	(0.88)(0.99)(0.99)	0.863	5
RS-FS-FT	(0.94)(0.93)	0.874	4
RS-FS-ES-ET-FT	(0.94)(0.98)(0.99)(0.99)	0.903	1

Table 1. *Pathways through the graph of Figure 2, with the edge probabilities and the total probability along the path (all of the edge probabilities multiplied, i.e., the assumption of independence). The final column is the rank by accuracy of the path (highest probability). In this case, the longest path, RS-FS-ES-ET-FT, has the highest predicted accuracy, of 0.903.*

In Table 1, we find that the longest path—that is, the path with the most operations—has the highest expected accuracy, even though it has at least one more step than any other path. This is because each operation in this path, RS-FS-ES-ET-FT, has high (0.94) or very high (0.98-0.99) accuracy. The overall expected accuracy is 0.903, compared to 0.893 for the much simpler path RS-RT-FT.

Validation

In the translation system described above, English was determined to the be the "central" language of a multi-language system. This was determined heuristically, when it was noticed that the use of English as an intermediate language led to improved overall system accuracy, even when one or two extra operations are required.

In addition to the heuristic approach, a non-heuristic means of assessing the central language in a multi-language system can be performed. In Table 2, the matrix of error rates between languages in an operation is given. This matrix is for text-to-text translation for the EFIGS (English, French, Italian, German, Spanish) languages for the translation engine used. The translation accuracy from English to French, for example, is 0.99. The opposite direction, that of translating from French to English, has a similar but lower accuracy of 0.97.

Source \ To	English	French	Italian	German	Spanish
English	N/A	0.99	0.98	0.99	0.98
French	0.97	N/A	0.96	0.95	0.96
Italian	0.95	0.96	N/A	0.96	0.97
German	0.96	0.94	0.95	N/A	0.92
Spanish	0.95	0.97	0.98	0.93	N/A

***Table 2**. Translation accuracies from the source language (first column) to the destination languages (columns 2-6).*

The rows and columns of Table 2 provide insight into the working of the overall multi-language system. From the mean of the rows, for example, we see that when English is the source text, the mean translation accuracy is 0.985. For the other languages, this is substantially lower: 0.96 from French, 0.96 from Italian, 0.943 from German, and 0.958 from Italian. These indicate English is the most accurate, German the least accurate, with the three Latin languages intermediate in accuracy. Taking the mean of the columns, translations into English have a mean accuracy of 0.958. The means for translating into French, Italian, German, and Spanish are 0.952, 0.954, 0.946, and 0.946, respectively. While English is again the highest mean accuracy for the columns, the differences between the columns are less than a third of the differences between the rows. Thus, the differentiating accuracy for the entire system is the accuracy of English text into the other four languages. The lower accuracy of translating German text into the other four languages is also a characteristic of the system. Combined, these results indicate that English is the **central language** for the system (consistent with the results for the principal translation problem of this paper), and that German is probably the language for which the system builders had the least expertise. However, since French, Italian, and Spanish are more closely related in syntax and vocabulary, it is possible that this similarity collectively lifts their results above those of German, and that their proficiency in English is the most defensible finding.

In order to address whether or not the system has linguistic asymmetry, the ratios of "To/From" are computed for each language pair. For English and French, then, the "To/From" ratio for English is 0.99/0.97 = 1.021. For French, it is the inverse, 0.97/0.99 = 0.980. These ratios are collected in Table 3. As in Table 2, this table contains 20 relevant values (the diagonal is "not applicable", or N/A).

Source	English	French	Italian	German	Spanish
	To/From	To/From	To/From	To/From	To/From
English	N/A	1.021	1.032	1.031	1.032
French	0.980	N/A	1.000	1.011	0.990
Italian	0.969	1.000	N/A	1.011	0.990
German	0.970	0.989	0.990	N/A	0.989
Spanish	0.969	1.010	1.010	1.011	N/A

***Table 3**. Translation accuracy ratios to/from the other languages (ratios of To and From data in Table 2), computed to determine if translation asymmetries exist.*

The rows of Table 3 are analyzed using a simple z-value, z=|μ-1.0|/(σ/sqrt(n)). Here, n=4 since there are four values for each row. The mean of the row, μ, is the mean of the four non-diagonal values, and the standard deviation of these fours values is σ. The p-value (two-tailed) of the z-scores are shown, along with μ, σ, and the z-score in Table 4.

Language	Mean (μ)	Std (σ)	z-value	p(z-value)
	To/From	To/From		
English	1.029	0.005	10.83	0.000
French	0.995	0.013	-0.71	0.475
Italian	0.993	0.018	-0.84	0.401
German	0.985	0.010	-3.20	0.00136
Spanish	1	0.021	0	1.000

***Table 4**. Calculation of asymmetry. The z-value is computed, and the p-value is calculated from a z-table (two-tailed test). If p<0.05, then the language is considered asymmetric. In this table, English is positively asymmetric while German is negatively asymmetric.*

The results of Table 4 illustrate the asymmetric behavior of both English and German languages. English as a language translated to another language has higher accuracy than English as a language translated to from another language. German has the opposite behavior. This asymmetric behavior is a form of sensitivity analysis for the linguistic system. Any such asymmetries are indicative of overall system immaturity, meaning that there is room for improvement in the overall system accuracy, if just the right algorithm, training set, or meta-algorithm could be employed. Thus, the possibility of linguistic asymmetries should always be investigated. In the current system, however, it means that given the choice for a pipeline, we would prefer to move from English text and to German text as steps in a pipeline. This is because these steps have asymmetrically higher accuracy than their opposites, moving to English text and from German text.

Discussion

The approach outlined in the preceding Validation section is concerned with translation, but it could also be used for any other multi-stage text analytics process, including one extending from key words to summaries to documents to clusters of documents. The central analytic will be the one with the highest accuracy, and asymmetries in the steps between two types of data allow us to determine preferential elements in our processing pathways. Since we did not perform the Validation step on Russian language (due to insufficient training sets and proficiency), we cannot determine whether our Russian speech to French text translation benefitted from asymmetry (especially since the English-language pipeline was both translated into and out of). However, it is likely. Our findings for Tables 2, 3, and 4 show that translating out of English text into another language was a differentially accurate operation. Since this operation occurs (English text to French text), and no translation into English text from another language's text occurs, our surprisingly long pipeline, that is RS-FS-ES-ET-FT, likely does benefit from the same asymmetry outlined in the Validation section.

There are a number of concerns, or at least caveats, about the investigation presented. The first is that the different operations are not likely to be independent of each other. For example, if a sample of Russian speech is difficult to translate into French speech, it is likely that it is also difficult to translate into English speech, and maybe even to transform into Russian text. Thus, several of the probabilities listed on the edges may be correlated (the document set size was too small for conclusion interpretation of the existence of, or lack of existence of, independence among the edge operations). The second concern with the example is that there is no penalty for the number of operations performed: RS-FS-ES-ET-FT has four steps, while RS-RT-FT has only two steps. This could mean, for example, that RS-FS-ES-ET-FT costs twice as much to perform as RS-RT-FT, it takes twice as long to perform (since it requires twice the operations), and/or is much more sensitive to changes in the inputs (and is thus less robust to data drifting). At the time this investigation was performed, the linguistic software (Nuance and Speechworks) were for-charge services, meaning that the number of operations was correlated with cost.

An interesting finding, and one that is likely to be repeatable in other applications, is that having the analysis path pass through either English speech (ES) or English text (ET) nodes may have some particular advantages. Since English, based on the accuracies reported, is the "central" language for the overall system of text analytics operations, having the input content internalized as English may be highly advantageous for the repurposing of the content. Suppose that another language (e.g. Spanish or Mandarin Chinese) or another application (e.g. summarization or document clustering) is added to the system. Having the ES and ET information also allows the data analyst concerned with testing and configuration to have a lingua franca, as it were, for comparing two different systems. If every major text analytics task to be performed is channeled through ET and ES, then the ET and ES data sets can be "fairly" compared to one another for selecting an optimum system configuration. That is, the ET and ES "central" content is what can be used for benchmarking one system configuration versus another.

Irrespective of its overall advantages, the method shown in this section can be used to compare and contrast different pathways for multi-step text analytics tasks. Here, the function being evaluated is the optimum pathway for an important systems metric such as accuracy. The results show that software in which there is differential proficiency in one language – our so-named "central" language which was English – may benefit (in terms of accuracy, etc.) from pathways using this centrality, even if these pathways are lengthier than other pathways.

The benefit of this work to the cultural heritage sector resides largely in the ability to provide the most accurate translation possible when there is insufficient language expertise, training data, time, money, or other factors available for a human-directed translation. Analogous to historical documents in the optical character recognition space, generalized language translation often suffers from the "can't get there from here" problem wherein specific speech-to-speech, text-to-speech, and/or speech-to-text technologies do not exist. This is especially true for marginal and extinct languages, precisely ones that may be of particular interest to the library, archival, and museum communities. The research performed for this work demonstrate that circuitous pathways between a starting language and media (speech, text) and ending language and media do not necessitate lower accuracy. In fact, if a pipeline of transformations, such as those involving the English language in the example here, with differentially high accuracy can be found, the generalized translation path can be supported by early on-ramp into the native language of the recognition/translation software and late off-ramp from the same.

Acknowledgments

Steven Simske' s work on this paper was partially supported by NSF EAGER grant with the abstract number 1842577, Advanced Analytics, Intelligence and Processes for Disrupting Operations of Illicit Supply Networks.

References

[1] Oliver W. Morgan, Lisa Page, Sarah Forrester, Helen Maguire, "Polonium-210 Poisoning in London: Hypochondriasis and Public Health," Prehosp. Disaster Med., 23, 96 (2008).

[2] Steven J. Simske, Marie Vans, Functional Applications of Text Analytics Systems (River Publishers, 2021), pg. 237.

[3] Rafael D. Lins, Rafael Ferreira, Rinaldo Lima, Gabriel de Franca Pereira e Silva, Hilario Oliveira, Luciano Cabral, Jamilson Batista, Bruno Tenorio, Diego A. Salcedo, Steven J. Simske, "The CNN-

Corpus in Spanish: a Large Corpus for Extractive Text Summarization in the Spanish Language," Proceedings of the 2019 ACM Symposium on Document Engineering (DocEng '19), (2019), pgs. 1-4.

[4] Rafael Dueire Lins, Hilario Oliveira, Luciano Cabral, Jamilson Batista, Bruno Tenorio, Rafael Ferreira, Rinaldo Lima, Gabriel de Franca Pereira e Silva, Steven J. Simske, "The CNN-Corpus: A Large Textual Corpus for Single Document Extractive Summarization," Proceedings of the 2019 ACM Symposium on Document Engineering (DocEng '19), (2019), pgs. 201–210.

Author Biographies

Steven Simske is a Professor of Systems Engineering at Colorado State University. was at HP from 1994-2018, and was an HP Fellow, Vice President, and Director in HP Labs. He is currently the author of more than 450 publications and more than 200 US patents). He is an IS&T Fellow, and its immediate past President (2017-2019). Please refer to https://www.engr.colostate.edu/se/steve-simske/.

Marie Vans is a senior research scientist in HP Inc's CTO office in Fort Collins, currently focused on building large-scale data systems using cloud-based technologies. She is also an adjunct faculty member of San José State University's iSchool where she teaches a course called Design for Teaching & Learning in Social Virtual Reality.