Semantic Overlap Computation

Kanishka Misra, Hemanth Devarapalli

1/22/2019

Learner Errors in English

In this example, we consider an error annotated corpus of 1244 short essays, each written by a unique learner of english belonging to one of 16 different first languages (L1).

Since we are covering semantic errors, we extract the errors made in content words (adjectives (J), nouns (N), verbs (V) and adverbs (Y)).

Let's look at an example by a learner whose L1 is Russian:

While this example contains many other errors, we only focus on one for this analysis. The error word is in **bold**.

Original Answer:

Dear Jane Clark, I am writing to you to give some opinions of mine abut The International Arts Festival. It was a great idea and I had the pleasure of resting and relaxing, plus of ***getting*** *some important knowledge from it. On the other hand there were some disadvantages: 1. Stars and artists wee only from six countries 2. Some concert halls were too small. So it was impossible to get tickets there. 3. I think that there were not enough plays and films. To sum up I want to tell you bout my suggestions for next year's festival. It would be wonderful to buy some books or programms with signatures of all the stars and artists taking part in the festival. Also I think that dance shows should include different styles of dancing (e.g. national dancing of all the countries) to impress audience and not to bore with the similar scenes . I hope this letter will help you in organying Yours faithfully*

Correct Answer:

Dear Jane Clark, I am writing to you to give you some opinions of mine about The International Arts Festival. It was a great idea and I had the pleasure of resting and relaxing, plus of ***acquiring*** *some important knowledge from it. On the other hand there were some disadvantages: 1. The stars and artists were from only six countries . 2. Some concert halls were too small. So it was impossible to get tickets there. 3. I don't think that there were enough plays and films. To sum up I want to tell you about my suggestions for next year's festival. It would be wonderful to buy some books or programmes with the signatures of all the stars and artists taking part in the festival. Also I think that the dance shows should include different styles of dancing (e.g. the national dance of all the countries) to impress the audience and not to bore with the similar scenes . I hope this letter will help you with organizing the festival . Yours sincerely*

Analysis

From the above short essay responses, we get the incorrec and replacement annotations

Incorrect word in English: **getting** Replacement as annotated: **acquiring**

We then take this (incorrect, correct) word pair and translate it into the person's L1 language using the Microsoft Azure Text translation API [1]. We use this API because unlike Google's Cloud API for translations,

[1] https://docs.microsoft.com/en-us/azure/cognitive-services/Translator/reference/v3-0-reference

it provides us with alternate pronunciations which prove beneficial while querying for words that can be expressed with different parts of speech.

Translated form: **получение** Translated replacement: **приобретения**

Semantic Overlap

We then use the fasttext english and russian word vectors (300 dimensions) to calculate the 10 nearest neighbors for the *(incorrect, correct)* pair in their respective language vector spaces.

The 10 nearest neighbors for each of these words are shown in table 1.

Table 1: 10 nearest neighbors of the words in the given example.

English		Russian	
getting	**acquiring**	**получение**	**приобретения**
gettting	Acquiring	Получение	покупки
gettng	reacquiring	предоставление	приобритения
geting	obtaining	-получение	продажи
get	acquire	приобретение	получения
gettign	re-acquiring	неполучение	приобретении
gettiing	acquired	получения	приобретение
gettig	procuring	лучение	перепродажи
got	acqruing	наполучение	преобретения
getitng	acquisition	получении	приобретению
gotten	owning	получение	использования

To compute the semantic overlap between **i** and **c** in a given language, we take each word and compute its partial overlap to the other word. The partial overlap of a word x with word y is computed as:

$$PO(x, y) = \frac{1}{k} \sum_{y' \in NN_k(y)} cos(x, y')$$

Where k is the number of nearest neighbors, here 10 and $NN_k(y)$ is the nearest neighbor function to print the k nearest neighbors of the word y based on cosine similarity. Thus, we get the partial overlap of **i** with **c** and **c** with **i**. We then compute the Semantic Error Overlap, SEO by taking the mean of the two partial overlaps $PO(i, c)$ and $PO(c, i)$.

Based on the example:

For **getting** and **acquiring**,

We compute $PO(getting, acquiring)$ and $PO(acquiring, getting)$ as:

`getting with [Acquiring, reacquiring, obtaining, acquire, re-acquiring, acquired, procuring, acqruing, acquisition, owning] = 0.32476866`

and

`acquiring with [gettting, gettng, geting, get, gettign, gettiing, gettig, got, getitng, gotten] = 0.27844474`

Taking the average, we have: $SEO(getting, acquiring)$ ~ 0.302

Similarly, we compute *PO(получение, приобретения)* = 0.41347638 and *PO(приобретения, получение)* = 0.41594142 to get *SEO(получение, приобретения)* ~ 0.415