RE: automatically collecting psycholinguistic features of Swedish words from corpora?

Paridon, Jeroen van

Mon 4/15/2019 12:17 PM

To: Montero-Melis, Guillermo <Guillermo.MonteroMelis@mpi.nl>;

2 attachments (807 KB)

nl_results.tsv; en_results.tsv;

Hi Guillermo,

I haven't been able to do all I'd like to do on this problem this weekend, but here's what I *have* gotten done.

I've attached two files, one for Dutch and one for English. (Open the files in Excel for optimal viewing enjoyment.)
These lists were compiled using the following method, roughly speaking:

1. Retrieve 300-dimensional word vectors for the words in the list
2. Compute the median across all words for every dimension
3. Compute the cosine distance to the median vector for every word vector we have available
4. Sort these words by cosine distance
5. Store the top 10,000 words in a file (together with their cosine distance to the median vector, and whether they are part of the original word list)

Some notes:

- This seems to have worked reasonably well. For Dutch, the words close to the median seem to be what you'd expect (walking, hopping, running), for English the words close to the median are a little weirder but still clearly motion-related (saunter, scurry, careen); this may indicate our method isn't perfect, but it's also possible the original English word list is just a little biased towards more unusual motion verbs.
- The cosine to median method is the most simple, straightforward method I could come up with, but there are of course other ideas we could go with, for instance:
  - Cosine to mean vector (I've tried this, doesn't seem to make too much of a difference)
  - Cosine to a particular word or set of words (i.e., not the complete word list, but a specific selection from it; this could mitigate the outsize impact unusual motion verbs seem to have on the English median vector)
  - Some slightly more sophisticated statistical method (like PCA) to extract the most motion-relevant dimensions from the vectors, and scoring words based on those (I think this is a promising idea, but it would just be more work to implement)

Let me know what you think.

Kind regards,

Jeroen

From: Montero-Melis, Guillermo <Guillermo.MonteroMelis@mpi.nl>
Sent: vrijdag 12 april 2019 2:59 PM
To: Paridon, Jeroen van <Jeroen.vanParidon@mpi.nl>
Subject: Re: automatically collecting psycholinguistic features of Swedish words from corpora?

hi Jeroen,
Hope your meetings went well!

Here I am attaching the Swedish verb list.

So let me summarize and clarify, to be on the safe side:

The tasks:
1) For the Dutch and English lists (previous email), we're interested in looking for other manner verbs by looking at words (verbs) that have low cosine similarities to the current set. One of the Dutch files contains a subset of the verbs in the other file, which is a simple way to validate the technique by checking whether we can retrieve the ones we left out
2) For Swedish, you will try to retrieve the following information: Word frequency, Lemma frequency (?), Bigram frequency, Trigram frequency, Grammatical ambiguity (?)

Priorities:
- Task 1 takes precedence over task 2
- For task 1, Dutch takes precedence over English

Once you have the new retrieved verbs from Task 1, send them along so I can share them with our RA for coding.

Cheers,
Guillermo

From: Montero-Melis, Guillermo
Sent: Friday, April 12, 2019 9:28:38 AM
To: Paridon, Jeroen van
Subject: Re: automatically collecting psycholinguistic features of Swedish words from corpora?

I'm coming over. Several verb lists attached

From: Montero-Melis, Guillermo
Sent: Thursday, April 11, 2019 11:19:36 PM
To: Paridon, Jeroen van
Cc: Margareta Majchrowska
Subject: Re: automatically collecting psycholinguistic features of Swedish words from corpora?

Hi Jeroen,
9:30 tmw works perfectly! I'll drop by your office then.
cheers
g

From: Paridon, Jeroen van
Sent: Thursday, April 11, 2019 11:08:33 PM
To: Montero-Melis, Guillermo
Cc: Margareta Majchrowska
Subject: Re: automatically collecting psycholinguistic features of Swedish words from corpora?

If you're going to be out of the office for a while, maybe we should meet tomorrow before the lab meeting. I can make sure to come in at 9:30, if that works for you.
I'm happy to help with points 3-7, of course. With regards to collecting the 8-13 norms in Swedish: any data is of course better than nothing.

Jeroen

From: Montero-Melis, Guillermo
Sent: Thursday, April 11, 2019 10:57:27 PM
To: Paridon, Jeroen van
Cc: Margareta Majchrowska
Subject: Re: automatically collecting psycholinguistic features of Swedish words from corpora?

Hi Jeroen,
Thanks, that's very useful!

So here are my thoughts:
- If 3-6, and possibly 7, are easy for you, it'd be great if you could help us with that.
- By 7 my guess is that they indeed mean that kind of noun/verb/adjective ambiguity; they don't specify it, but it's the only thing that makes sense given the context (they checked these things somewhere, god knows where, so it has to be information that is more or less easily available for English)
- For 8-13, we'll go on with our norming idea then, since I am doubtful I will find any of these norms for Swedish. We can be pragmatic about it and collect norms from 20 people or so. For this, I think (hope) it will suffice.

Tmw we have our lab meeting 10-12. I'll leave early in the afternoon and next week I won't come in. So it'd have to be the week after. Are you going to be around for the last 3rd of April / beginning of May?

Thanks again!

Best,
Guillermo

From: Paridon, Jeroen van
Sent: Thursday, April 11, 2019 8:27 PM
To: Montero-Melis, Guillermo
Cc: Margareta Majchrowska
Subject: Re: automatically collecting psycholinguistic features of Swedish words from corpora?

Hi Guillermo,

3, 5, and 6 would be trivial for me to retrieve if you send me a list of words. 4 is likely trivial as well, although I'd need to check if the lemmatized corpus I have for Swedish is large enough to get reliable numbers. For 7, do you mean grammatical ambiguity in the sense that a word like "lift" can be both verb and noun, or do you mean phrase-length (or longer) units that can be interpreted multiple ways? (The former I can probably find out for you, the latter would be next to impossible.)

8 through 13 are not trivial, unfortunately. There is a method for matching semantic dimensions in different languages, but it's computationally intensive and very much a work in progress (one that I won't get to work on very much until after I finish my thesis). If you can find any of these norms in Swedish (even if it's for entirely different words) that would make it easier because then I can just extend the norms within the language, instead of having to try to predict across languages.

Either way, we can discuss this in person if that's helpful. Tomorrow I'll be in my office between 10:00 and 11:45 for certain; outside those hours I might be in meetings.

Kind regards,

Jeroen

From: Montero-Melis, Guillermo
Sent: Thursday, April 11, 2019 5:36:12 PM
To: Paridon, Jeroen van
Cc: Margareta Majchrowska
Subject: automatically collecting psycholinguistic features of Swedish words from corpora?

Hi Jeroen,
(cc-ing Margareta, a Master's student in Stockholm who is working with me on this project)

Let me ask you a question following up on something you mentioned the other day during our meeting. I want to know how difficult / time demanding it would be for you to do your magic using your vector space tools to solve an issue of stimuli validation. Let me explain:

For a replication study, we were planning to collect some norms for our Swedish materials (the original study was conducted in English; I am attaching the study we are replicating for reference, Shebani and Pulvermüller, 2013, *Cortex*; **S&P** for short)

We would like to match two lists of words along the following variables (cf. p.225, Table 1 in S&P):

1. Number of letters
2. Number of phonemes
3. Word frequency
4. Lemma frequency
5. Bigram frequency
6. Trigram frequency
7. Grammatical ambiguity
8. Valence
9. Arousal
10. Imageability
11. Visual relatedness
12. Body relatedness
13. Action relatedness

Unfortunately, S&P don't cite any sources about where they took this information from. In any case: Some of these are trivial (1) or foreseeably easy to obtain even for Swedish (2-4). Others are slightly more tricky, at least for me (5-7). For 8-10 it's easy to find English norms (Brysbaert's stuff). For 11-13 we've struggled a bit more to figure out who collected them, and thus we are not entirely sure what the numbers mean (probably ratings on Likert scales, but on which exactly? Anyway, that's a separate issue).

My approach right now was to collect data for 8-13 ourselves: set up an online survey form and have 20 native speakers or so rate our critical words along these dimensions (probably interspersing them with other random words to not bias the ratings). But I just talked to Markus and he told me you might be able to get some of these measures for Swedish using a computational approach? Is it the case? And if so, how much work would it take? What information would you need to do that? E.g., for the Visual Relatedness ratings, suppose we find a study that has collected such norms (à la Brysbaert); would that be enough to get a measure of this for Swedish as well? I guess it'd help me to know how you would go about...

Perhaps it's easier to discuss this in person, so let me know if you have a minute and I can drop by your office.

Cheers,
Guillermo