OE Taalkunde, Dag van het Onderzoek, 2014
Corpus-based research – Analysing data

Dirk Speelman[1]

[1] representing QLVL, KU Leuven

Overview

Introduction

Technique 1: regression analysis

Technique 2: distance metrics and scaling (e.g. MDS)

Technique 3: collocations and keywords

Technique 4: vector space models (VSMs)

Overview

Introduction

Technique 1: regression analysis

Technique 2: distance metrics and scaling (e.g. MDS)

Technique 3: collocations and keywords

Technique 4: vector space models (VSMs)

Introduction

- The emphasis in this talk is on the study of language usage

Introduction

- The emphasis in this talk is on the study of language usage
- We zoom in on patterns in usage

Introduction

- The emphasis in this talk is on the study of language usage
- We zoom in on patterns in usage
- Pattern detection implies quantitative analysis

Introduction

- The emphasis in this talk is on the study of language usage
- We zoom in on patterns in usage
- Pattern detection implies quantitative analysis

Hence we zoom in on quantitative analysis in this talk

Introduction

- The emphasis in this talk is on the study of language usage
- We zoom in on patterns in usage
- Pattern detection implies quantitative analysis

Hence we zoom in on quantitative analysis in this talk

By which we do not want to imply that qualitative analysis of texts isn't worthwhile

Introduction

We will discuss four techniques.

- Techniques one and two (regression analysis and multidimensional scaling) are general purpose statistical techniques that have many applications within and outside of corpus linguistics. Many individual researchers will benefit from familiarizing themselves with these techniques.
- For these techniques we'll take a more hands-on approach and we'll show the techniques at work in R (although the R session will only cover 'highlights', not fully-fledged analyses).

Introduction

- Techniques three and four (collocations/keywords and vector space models) are quantitative computational techniques specific for corpus linguistics. These techniques are typically either fully automated or only conducted by a relatively small number of specialized research teams. Most individual researchers will not need to handle the technical details of these techniques themselves.
- These techniques will be discussed at a more conceptual level.

Introduction

Why R?

- large user community: stable, well documented tool
- easy integration with retrieval (decent support for string manipulation)
- some good textbook on statistics for linguistics in R (e.g. Baayen, 2008).

The R session we'll use today can be found at `http://wwwling.arts.kuleuven.be/qlvl/R-demos/dag-onderzoek-2014.html`.

Overview

Introduction

Technique 1: regression analysis

Technique 2: distance metrics and scaling (e.g. MDS)

Technique 3: collocations and keywords

Technique 4: vector space models (VSMs)

Regression analysis

Research questions

- Can we statistically model the probabilistic behavior of a response variable (= the phenomenon of interest) on the basis of a number of predictors (= candidate explanatory variables)?
- What is the impact of some predictor X on the response variable, taking into account that there are also a number of other predictors at work simultaneously?
- What is the combined impact of a number of predictors on the response variable?
- Are there interactions between predictors in the model (meaning that the impact of the one predictor on the response variable is different, depending on the value of the other predictor)?

Regression analysis

- Regression analysis is part of inferential statistics. We test for significance, which means that we test if we are sufficiently confident that the patterns we observe have counterparts in the population from which we study a sample.
- But there are some caveats:
  - In corpora (which are observational data) we can only observe correlations; we cannot prove causal relations.
  - Regression analysis assumes that our samples are drawn at random from the population (or are constructed in some other way that precludes selection bias and inter-item dependence). However, corpora are never perfectly random samples from some population. This is something we'll have to take into account. We need many replication studies. We also have to be very careful for inter-item dependence (e.g. by taking into account random effects).

Logistic regression analysis

- Logistic regression analysis is a type of regression analysis in which the response variable is categorical.
- We'll discuss the case of binary logistic regression analysis, in which the response variable has two levels (i.e. two possible values).

Logistic regression analysis

Case study: `http://wwwling.arts.kuleuven.be/qlvl/R-demos/dag-onderzoek-2014.html`

Overview

Introduction

Technique 1: regression analysis

Technique 2: distance metrics and scaling (e.g. MDS)

Technique 3: collocations and keywords

Technique 4: vector space models (VSMs)

Multidimensional Scaling (MDS)

Research questions

Can we represent the dissimilarities between the items in our data in a low-dimensional space in such a way that the distances in low-dimensional space reflect these dissimilarities reasonably well? [in the specific technique that we'll discuss, isoMDS, this means a 'stress' of not much more that about 15%.]

In metric MDS the relation between the distances and the dissimilarities must be linear. In non-metric MDS the relation must simply be monotonic. We'll illustrate isoMDS, which is a kind of non-metric MDS.

Multidimensional Scaling

- isoMDS is an exploratory technique. Its main purpose is to help the research oversee complex data, and to help the researcher detect patterns. However, it doesn't allow the research to establish 'significance' of these patterns.
- In other words, it only helps te researcher to detect patterns in the sample.
- These patterns may inspire the researcher to formulate hypotheses (about the population) to be tested in follow-up research.

Multidimensional scaling

Case study: `http://wwwling.arts.kuleuven.be/qlvl/R-demos/dag-onderzoek-2014.html`

CGN components

a. Spontaneous conversations ('face-to-face')
b. Interviews with teachers of Dutch
c. Spontaneous telephone dialogues (recorded via a switchboard)
d. Spontaneous telephone dialogues (recorded on MD via a local interface)
e. Simulated business negotiations
f. Interviews/discussions/debates (broadcast)
g. (political) Discussions/debates/meetings (non-broadcast)
h. Lessons recorded in the classroom
i. Live (eg sports) commentaries (broadcast)
j. Newsreports/reportages (broadcast)
k. News (broadcast)
l. Commentaries/columns/reviews (broadcast)
m. Ceremonious speeches/sermons
n. Lectures/seminars
o. Read speech

Profiles

de, den
een, ne, nen
geen, gene, genen
iedere, iederen
elke, elken
deze, dezen
die, dieje, diejen, dien, diene, dienen
mijn, mijne, mijnen
je, jouw
zijn, zijne, zijnen
haar, hare, haren
onze, onzen
uw, une, uwen
hun, hunne, hunnen
...

Overview

Introduction

Technique 1: regression analysis

Technique 2: distance metrics and scaling (e.g. MDS)

Technique 3: collocations and keywords

Technique 4: vector space models (VSMs)

Collocation analysis

Research question

Which empirical co-occurrence patterns (between words) are characteristic of my data?

Collocation analysis

Research question

Which empirical co-occurrence patterns (between words) are characteristic of my data?

It is important to distinguish between:

- empirical co-occurrence pattern, which we'll call collocations
- (the more theoretical concept of) lexicalised word combinations, which we'll call multiword expressions

Collocation analysis

Research question

Which empirical co-occurrence patterns (between words) are characteristic of my data?

It is important to distinguish between:

- empirical co-occurrence pattern, which we'll call collocations
- (the more theoretical concept of) lexicalised word combinations, which we'll call multiword expressions

We can only directly detect the former in corpora, not the latter.

Collocation analysis

- In practice, collocation analysis has been used for detecting the following related (and often overlapping) but definitely not identical phenomena (the list is not exhaustive):
  - semantic nuances between related words, lexical restrictions, idioms, lexical collocations, terms (in the terminological sense), multiword expressions in the NLP sense, clichés, cultural stereotypes, semantic compatibility, stylistic and register compatibility, proper names, compounds, semantic relatedness between collocating words, ...
- These differences in research interest and different applications (NLP, lexicography, ...) have contributed to the diversity in the tests and measures that are being used in collocation analysis.

Collocation analysis

Evert (2008) uses the following terminology to distinguish between three different perspectives on (an hence three different procedures to establish) co-occurrence.

- surface co-occurrence is the co-occurrence within the same 'span' (or 'window') in a text. A window is defined as 'at most so many words to the left and so many words to the right of the node', where we treat one of the co-occurring words as the 'node' (or 'target word') and the other one as the 'collocate'. E.g. (L2,R3) means a span of two words to the left and three to the right. Collocations at the surface level are called surface collocations.

Collocation analysis

- textual co-occurrence is the co-occurrence within the same textual unit. The textual unit that is used can vary between analyses. It can either be a subclause, a clause, a sentence, a paragraph or even a complete document. Collocations at the textual level are called textual collocations.
- syntactic co-occurrence finally is the co-occurrence in two (directly or indirectly linked) roles in the same instance of a syntactic pattern. For instance the combination of adjective and noun, or the combination of verb and (head of the) direct object, etc. Collocations at the syntactic level are called syntactic collocations.

Collocation analysis

What do you need?

An unannotated corpus will do for surface collocation analysis and some types of textual collocation analysis at the level of wordforms, but

- obviously lemma and POS information are needed for any type of collocation detection at the level of lemmata or for disambiguating words on the basis of their POS code.
- parse trees are needed for some types of textual collocation analysis and for syntactic collocation analysis.

Collocation analysis

Which software?

- stand-alone concordancers (AntConc, WordSmith Tools, ...)
- web corpus query systems (BYU web corpora, Sketch Engine, CQPweb, ...)
- R (possible via R's decent text manipulation functions, but requires some scripting)
- Python and other scripting languages (excellent text manipulation functions)
- ...

Collocation analysis

What do the results look like?

- You typically obtain a list of collocations, ranked according to some measure of association strength, with the strongest associations on top.
- Researchers then typically treat the top of the list as the (most) important collocations; they either use some threshold value of association strength to make a categorical selection or they simply select the top n (e.g. top 50) of the list.

Collocation analysis

Effect size measures of association strength

- Effect size measures try to capture the magnitude of the effect, but are insensitive to how much evidence there is that the effect is a real property of the population, and not merely the result of random variability in the sample.
- Therefore effect size measures become less reliable if they are based on low frequencies.

Collocation analysis

Some examples of effect size measures

- PMI, pointwise mutual information, is designed to capture how much information on the probability of the presence of the collocate is available in the knowledge that a given node is present. [The measure is symmetric and therefore also captures how much information on the probability of the presence of the node is available in the knowledge that a given collocate is present.]

Collocation analysis

Some examples of effect size measures

- The DICE coefficient is useful to identify fixed expressions. It only reaches high values if the following two situations both apply
  - given the presence of the node, it is very likely that the collocate will also be present
  - given the presence of the collocate, it is very likely that the node will also be present

Collocation analysis

Some examples of effect size measures

- The Odds Ratio expresses how different the odds of the collocate are, depending on whether the node is or isn't present [and since this is a symmetrical measure, it also expresses how different the odds of the node are, depending on whether the collocate is or isn't present].

Collocation analysis

Some examples of effect size measures

- Delta P is an asymmetrical measure:
  - On the one hand, Node-to-Collocate Delta P expresses the difference of proportions of the collocate given the presence and absence of the node respectively. It expresses to which extent the node 'selects' the collocate.
  - Conversely, Collocate-to-Node Delta P expresses the difference of proportions of the node given the presence and absence of the collocate respectively. It expresses to which extent the collocate 'selects' the node.

Collocation analysis

Measures based on inferential statistical tests

- Measures based on inferential statistical tests capture how much evidence there is in our sample that the effect we observe is a real property of the population, and not merely the result of random variability in the sample. The actual measure is either the p-value or the test statistic of the statistical test.
- These measures in a sense 'favor' high frequency words, because high frequency implies more evidence. Effects in very high frequency words can obtain 'significance', even if the effect size is modest.

Collocation analysis

Some examples of measures based on inferential tests

- The test statistics of the χ^2 squared test or of the G^2 test. These are not to be used with very low frequency words.
- The p-value of the Fisher Exact test (smaller p-value means stronger association), which can also be used with very low frequency words.
- The test statistics of a t-test, which from a mathematical point of view is unsafe to use in this context (because of the possible violation of assumptions that underly the test), but is reported to yield intuitively plausible results.

Keyword analysis

Research question

Which words are empirically associated with our target corpus, when compared to our reference corpus?

Keyword analysis

Research question

Which words are empirically associated with our target corpus, when compared to our reference corpus?

It is important to understand that keyword analysis

- merely measures empirical association
- has no direct access to more conceptual characteristics of the keywords

Keyword analysis

- Just like collocation analysis, keyword analysis too is used for many different applications and types of research:
  - content analysis, term detection, study of register differences, study of stylistic variation, ...

Keyword analysis

What do you need?

- We need two corpora, a target corpus and a reference corpus. The second one should be big enough to help us establish what the reference behaviour of our words is like.
- Obviously lemma and POS information are needed for any type of keyword detection at the level of lemmata or for disambiguating words on the basis of their POS code.

Keyword analysis

Which software?

- stand-alone tools (AntConc, WordSmith Tools, ...)
- R (possible via R's decent text manipulation functions, but requires some scripting)
- Python and other scripting languages (excellent text manipulation functions)
- ...

Keyword analysis

What do the results look like?

- You typically obtain a list of keywords, ranked according to some measure of association strength, with the strongest associations on top.
- Researchers then typically treat the top of the list as the (most) important keywords; they either use some threshold value of association strength to make a categorical selection or they simply select the top n (e.g. top 50) of the list.

Keyword analysis

What do the results look like?

- You typically obtain a list of keywords, ranked according to some measure of association strength, with the strongest associations on top.
- Researchers then typically treat the top of the list as the (most) important keywords; they either use some threshold value of association strength to make a categorical selection or they simply select the top *n* (e.g. top 50) of the list.

In principle, all measures of association discussed in collocation analysis can also be used in keyword analysis.

Collocation analysis and keyword analysis

Pitfalls and some advice

- In most cases it is wise to look at at least one effect size measure and at least one inferential test based measure. Important collocations or keywords in your data should score (relatively) high on both scales.
- Don't treat p-values in these types of analyses the way you would in traditional statistical tests. You are making so many comparisons that their meaning has changed.
- Both keyword and collocation detection can be sensitive to topic bias and other (idiosyncratic) properties of texts and contexts. Measures that take into account dispersion and range information (not discussed here) will give more robust results.

Collocation analysis and keyword analysis

Some examples of collocation analysis and keyword analysis

- Study of (lexical) regional and register variation
- Term extraction
- Presence of collocation pattern as predictor in alternation studies

Overview

Introduction

Technique 1: regression analysis

Technique 2: distance metrics and scaling (e.g. MDS)

Technique 3: collocations and keywords

Technique 4: vector space models (VSMs)

Type-based vector space models

Research questions

What is the distance between words in terms of how they are used (with usage being operationalised as a wide range of 'features' of the contexts in which the words are used)?

Type-based vector space models

Underlying methodological questions are

- Can we use these distances as a proxy for semantic similarity or relatedness?
- Which operationalisation of distance (i.e. which parametrization of the procedure) can be used as a proxy for which type of semantic similarity or relatedness?

Type-based vector space models

Underlying methodological questions are

- Can we use these distances as a proxy for semantic similarity or relatedness?
- Which operationalisation of distance (i.e. which parametrization of the procedure) can be used as a proxy for which type of semantic similarity or relatedness?

More applied underlying questions are

- Can these distances help us improve the performance of NLP and information retrieval tasks (thesaurus building, identification of translation equivalents, question answering, ...)?
- Which operationalisation of distance is best for optimizing performance in which specific tasks?

Type-based vector space models

Step 1: item (= target word) by context feature frequency matrix

In bag-of-words vector space models context features are surface collocates. Frequencies in the cells are typically transformed into PMI values.

	home	*drink*	*traffic*	*wheel*	...
car	...	...	...	...	...
vehile	...	...	...	...	...
coffee	...	...	...	...	...
...	...	...	...	...	...

Type-based vector space models

Step 1: item (= target word) by context feature frequency matrix

In document based vector space models context features are the textual units we know from textual collocation. The use of the term 'documents' here is as flexible as the use of the term 'textual unit' was in collocation analysis. Frequencies in the cells are typically transformed into PMI values.

	doc 1	*doc 2*	*doc 3*	*doc 4*	...
car	...	...	...	...	...
vehile	...	...	...	...	...
coffee	...	...	...	...	...
...	...	...	...	...	...

Type-based vector space models

Step 1: item (= target word) by context feature frequency matrix

In dependency-based vector space models context features are syntactic collocates. Frequencies in the cells are typically transformed into PMI values.

	obj-of-drive	*obj-of-park*	*subj-of-crash*	*obj-of-drink*	...
car	...	...	...	...	...
vehile	...	...	...	...	...
coffee	...	...	...	...	...
...	...	...	...	...	...

Type-based vector space models

Step 2: item (= target word) by item similarity matrix

In all vector space models we then calculate similarities between rows (typically using the cosine of the angle between the row vectors as a similarity measure). This results in a similarity matrix:

	car	*vehicle*	*coffee*	...
car	1	...	...	...
vehile	...	1	...	...
coffee	...	...	1	...
...	...	...	...	1

Type-based vector space models

What do we need?

- We need a large corpus; for the methods to be informative, row vectors must not be too sparse
- An unannotated corpus will do for bag-of-words and some types of document-based VSMs at the level of wordforms
- obviously lemma and POS information are needed for any type of calculation at the level of lemmata or for disambiguating words on the basis of their POS code.
- parse trees are needed for some types of document-based VSMs and for dependency-based VSMs

Type-based vector space models

Which software?

- Although several general purpose distributional semantics software packages are available, most research groups working on vector space models work with their own code base.
- An important reason for this is that such research groups want to explore new types of parameter settings not (yet) supported in these general purpose tools.
- At QLVL we work with a combination of Python scripts for retrieval, MATLAB scripts for matrix calculations, and R (scripts) for additional statistical analysis. Our MATLAB scripts run on the KU Leuven HPC cluster. Our Python and R scripts run on the QLVL linux servers.

Type-based vector space models

What do the results look like?

- As we already saw, the main result is a huge word by word matrix with in its cells the similarities between the words.
- From this matrix one can retrieve more specific information, such as the k words that are closest to some target word (we speak of the k nearest neighbours of the target word in 'vector space').
- We could also try to visualise the information in the matrix, e.g. by turning the similarity matrix into a dissimilarity matrix, and by then applying MDS to this dissimilarity matrix.

Type-based vector space models

Some applications

- detection of regional lexical variation (when the nearest neighbour of a word in variety A is not the same word in variety B); a similar approach can be used for some other types of lectal variation.
- semi-automatic calculation of onomasiological profiles

Token-based vector space models

Research questions

- What is the distance between different tokens of the same word in terms of how they are used (with usage being operationalised as a wide range of 'features' of the contexts in which the tokens are used)? We say that these tokens together form a 'token cloud' in 'vector space'.
- What is the distance (or overlap) between 'token clouds' of different words in 'vector space'?
- ...

Token-based vector space models

Underlying methodological questions are

- Can we use these distances as a proxy for semantic similarity or relatedness?
- Which operationalisation of distance (i.e. which parametrization of the procedure) can be used as a proxy for which type of semantic similarity or relatedness?

Token-based vector space models

Underlying methodological questions are

- Can we use these distances as a proxy for semantic similarity or relatedness?
- Which operationalisation of distance (i.e. which parametrization of the procedure) can be used as a proxy for which type of semantic similarity or relatedness?

More applied underlying questions are

- Can these distances help us improve the performance of NLP and information retrieval tasks (word sense disambiguation, ...)?
- Which operationalisation of distance is best for optimizing performance in which specific tasks?

Token-based vector space models

Step 1: item (= token) by context feature frequency matrix

For instance, context features are surface collocates in that specific token. Of course cell frequencies will be very sparse.

	home	*drink*	*traffic*	*wheel*	...
car (token 1)	...	...	...	...	...
car (token 2)	...	...	...	...	...
car (token 3)	...	...	...	...	...
...	...	...	...	...	...

Token-based vector space models

Step 2: richer item (= token) by context feature frequency matrix

We make our sparse matrix richer by performing the following operation on each row

- we identify the column titles of all non-empty cells; these are our token-specific collocates and they will be the 'items' in the next step
- we retrieve the 'row vectors' of these 'items' in a (bag-of-words) type-based item by feature matrix (with PMIs in the cells)
- the new, richer row representation of the token will be the sum (or average) of the 'row vectors' from the previous step

Now we have a richer matrix.

Token-based vector space models

Step 3: token by token similarity matrix

We then calculate similarities between rows (typically using the cosine of the angle between the row vectors as a similarity measure). This results in a similarity matrix:

	car (token 1)	*car (token 2)*	*car (token 3)*	...
car (token 1)	1	...	...	...
car (token 2)	...	1	...	...
car (token 3)	...	...	1	...
...	...	...	...	1

Token-based vector space models

What do the results look like?

- As we already saw, the main result is a huge token by token matrix with in its cells the similarities between the tokens.
- From this matrix one can retrieve more specific information, such as the k tokens that are closest to some target token (we speak of the k nearest neighbours of the target token in 'vector space').
- We could also try to visualise the information in the matrix, e.g. by turning the similarity matrix into a dissimilarity matrix, and by then applying MDS to this dissimilarity matrix.

Token-based vector space models

Some applications

- detection of regional semasiological variation (when the token clouds of a word in two varieties do not overlap); the same technique can also be applied to some other types of lectal variation.

Thank you!

For more information:

dirk.speelman@arts.kuleuven.be