

Automatic Identification of Synonym Relations in the Dutch Parliament Thesaurus

LIS 2015 Colchester

Rosa Aga Christian Wartena Otto Lange
 Nelleke Aders

Hochschule Hannover, Abteilung Information und Kommunikation

Sept. 3rd, 2015

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions

Thesaurus Concepts

Concept

<http://www.tweedekamer.nl/thesaurus#1099>

Preferred term	Cultuurtechniek	<i>Agricultural engineering</i>
Other terms	afwatering bodemeigenschappen grondverbetering irrigatie ontginning teelttechniek	<i>drain</i> <i>soil properties</i> <i>soil improvement</i> <i>irrigation</i> <i>cultivation</i> <i>cultivation techniques</i>

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions



Overview

Motivation

Similarity of thesaurus terms

Experiment

Conclusions

Synonyms in the
Dutch Parliament
Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions

Overview

Motivation

Similarity of thesaurus terms

Experiment

Conclusions

Synonyms in the
Dutch Parliament
Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions



Information Processes in the Dutch Parliament

- ▶ All discussions in the parliament are accompanied by a large number of proposals, letters, reports, expertises, etc.
- ▶ All documents and the transcripts of the discussions are archived.
- ▶ Also additional relevant documents and news articles are collected and archived.
- ▶ All archived documents are provided with keywords from a controlled vocabulary.

Synonyms in the
Dutch Parliament
Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions

Information Processes in the Dutch Parliament

- ▶ All discussions in the parliament are accompanied by a large number of proposals, letters, reports, expertises, etc.
- ▶ All documents and the transcripts of the discussions are archived.
- ▶ Also additional relevant documents and news articles are collected and archived.
- ▶ All archived documents are provided with keywords from a controlled vocabulary.

Synonyms in the
Dutch Parliament
Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions



Synonyms in the Dutch Parliament Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions



Synonyms in the Dutch Parliament Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions

Thesaurus I

- ▶ Keywords are taken from the parliament thesaurus
- ▶ The parliament thesaurus was developed over the years by the *Dienst Informatievoorziening* of the the Dutch Parliament.
- ▶ About 4000 concepts.
- ▶ The concepts are organized in a loose hierarchy of broader/narrower and related terms.
- ▶ Each concept has one preferred term and on average also one non-preferred term.
- ▶ A number of non-preferred terms doesn't refer to a single concept but to a combination of two concepts.
- ▶ A number of concepts serve only the hierarchy building but cannot be used for annotation.




Thesaurus II

- ▶ The thesaurus is maintained in a proprietary system,
- ▶ but can be exported into a (proprietary) XML format.
- ▶ For this project we transformed this XML output to SKOS in RDF/XML.
- ▶ Transformation was done by an XSLT script

- parthes.rdf
 - Parlementsthesaurus
 - ALGEMEEN
 - BIRMA
 - CULTUUR
 - DEFENSIE
 - ECONOMIE
 - ENERGIE
 - FINANCIERING
 - FINANCIËN
 - GEZONDHEIDSZORG
 - GROTE OCEAANEILANDEN
 - GUADELOUPE
 - INDUSTRIE- EN DIENSTENSECTO
 - INFORMATIE EN COMMUNICATI
 - INTERNATIONALE POLITIEKE SIT
 - KUSTWATEREN
 - LANDBOUW**
 - AGRARISCHE SECTOR
 - AGRIFICATIE
 - ALTERNATIEVE LANDBOUW
 - CULTUURTECHNIEK**
 - LANDBOUWBEDRIJVEN
 - LANDBOUWGRONDEN
 - VOEDSELVOORZIENING
 - MAATSCHAPPIJ

CULTUURTECHNIEK

URI: <http://www.tweedekamer.nl/parlementsthesaurus#1099>

 **CULTUURTECHNIEK**, *afwatering, bodemeigenschappen, grondverbetering, irrigatie, ontginning, teelttechniek*

Broader: [LANDBOUW](#)

Narrower: [BEMESTING](#); [BESTRIJDINGSMIDDELEN](#); [GEWASBESCHERMING](#); [GEWASSCHADE](#)

Related: [BODEMDALINGEN](#); [WATERHUISHOUDING](#)

Keywords

- ▶ The thesaurus was changed over time.
- ▶ The rules for keyword assignments have changed over time.
- ▶ A uniform classification would be advantageous for retrieval
- ▶ There is not enough man power to annotate all documents with enough detail.
- ▶ Automatic or semi-automatic classification could help to overcome these problems.

Keywords

- ▶ The thesaurus was changed over time.
- ▶ The rules for keyword assignments have changed over time.
- ▶ A uniform classification would be advantageous for retrieval
- ▶ There is not enough man power to annotate all documents with enough detail.
- ▶ Automatic or semi-automatic classification could help to overcome these problems.

Automatic Classification

- ▶ In 2012 a commercial system for document classification was acquired.
- ▶ The system is based on a SVM trained with manually annotated documents from the Dutch Parliament.
- ▶ Results are disappointing
- ▶ Main problem: For most concepts there are not enough training documents
- ▶ **Keyword extraction is not classification!**

Synonyms in the
Dutch Parliament
Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions



Full Text Search

- ▶ Full text search is not possible for all (old) documents
- ▶ Full text search might have a low recall.
 - ▶ The search term is relevant but not mentioned in the text.
 - ▶ The vocabulary of official and governmental texts is quite different from daily language.
- ▶ Information specialists use thesaurus terms to search.
 - ▶ Most thesaurus terms are given in a form that doesn't occur in running texts.

Synonyms in the
Dutch Parliament
Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions

Thesaurus Enhancement

- ▶ Adding synonyms (as non-preferred terms) to the thesaurus concepts could help
 - ▶ for full text search (either used for query expansion or for document expansion)
 - ▶ help classification: concepts found by their new synonyms are added as additional features.
- ▶ Many concepts are stated in plural. (E.g. *Examens*, *Muziekscholen*, *Studenten* etc.) We can add singular forms as non-preferred term.
- ▶ **We can use other sources to add more synonyms to the concepts.**
- ▶ A method that is able to find new synonyms also should be able to distinguish pairs of synonyms from arbitrary pairs!
- ▶ Let us test a method!



Thesaurus Enhancement

- ▶ Adding synonyms (as non-preferred terms) to the thesaurus concepts could help
 - ▶ for full text search (either used for query expansion or for document expansion)
 - ▶ help classification: concepts found by their new synonyms are added as additional features.
- ▶ Many concepts are stated in plural. (E.g. *Examens*, *Muziekscholen*, *Studenten* etc.) We can add singular forms as non-preferred term.
- ▶ **We can use other sources to add more synonyms to the concepts.**
- ▶ **A method that is able to find new synonyms also should be able to distinguish pairs of synonyms from arbitrary pairs!**
- ▶ Let us test a method!



Overview

Motivation

Similarity of thesaurus terms

Experiment

Conclusions

Synonyms in the
Dutch Parliament
Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions

Distributional Similarity

- ▶ Two words are semantically similar if they occur in similar contexts.
 - ▶ Idea traces back to De Saussure, Wittgenstein, Harris
 - ▶ First implementations: Crouch (1990), Grefenstette (1992), Ruge (1992), Schütze & Pederson (1994)
- ▶ Recently a lot of progress
 - ▶ Compositionality
 - ▶ Overview studies comparing different approaches
- ▶ DS finds "semantically related" words, not only synonyms. What about co-labels for thesaurus concepts?



Word Contexts

- ▶ Syntagmatic Relations
 - ▶ E.g. selection relations; verbs with characteristic objects
 - ▶ Typical attributes: e.g. adjectives and nouns
- ▶ Paradigmatic Relations
 - ▶ Words that appear in similar contexts
 - ▶ e.g. words of the same syntactic class
 - ▶ Often: semantic similar words

Example

- ▶ *color* and *colour* will never co-occur
- ▶ *color* and *colour* will occur in similar contexts

Caution

- ▶ *Chronic* and *Disease* will often co-occur
- ▶ So, *Chronic* and *Disease* will occur in similar contexts!



Word Contexts

- ▶ Syntagmatic Relations
 - ▶ E.g. selection relations; verbs with characteristic objects
 - ▶ Typical attributes: e.g. adjectives and nouns
- ▶ Paradigmatic Relations
 - ▶ Words that appear in similar contexts
 - ▶ e.g. words of the same syntactic class
 - ▶ Often: semantic similar words

Example

- ▶ *color* and *colour* will never co-occur
- ▶ *color* and *colour* will occur in similar contexts

Caution

- ▶ *Chronic* and *Disease* will often co-occur
- ▶ So, *Chronic* and *Disease* will occur in similar contexts!



Word Contexts

- ▶ Syntagmatic Relations
 - ▶ E.g. selection relations; verbs with characteristic objects
 - ▶ Typical attributes: e.g. adjectives and nouns
- ▶ Paradigmatic Relations
 - ▶ Words that appear in similar contexts
 - ▶ e.g. words of the same syntactic class
 - ▶ Often: semantic similar words

Example

- ▶ *color* and *colour* will never co-occur
- ▶ *color* and *colour* will occur in similar contexts

Caution

- ▶ *Chronic* and *Disease* will often co-occur
- ▶ So, *Chronic* and *Disease* will occur in similar contexts!



Contexts for semantic similarity

- ▶ As a context we will use:
 - ▶ Two words to the right and left after removing stop words
 - ▶ Only open class words
 - ▶ Only words in a certain frequency range
 - ▶ Stems instead of words (from Treetagger)

Synonyms in the
Dutch Parliament
Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

Similarity of
thesaurus terms

Experiment

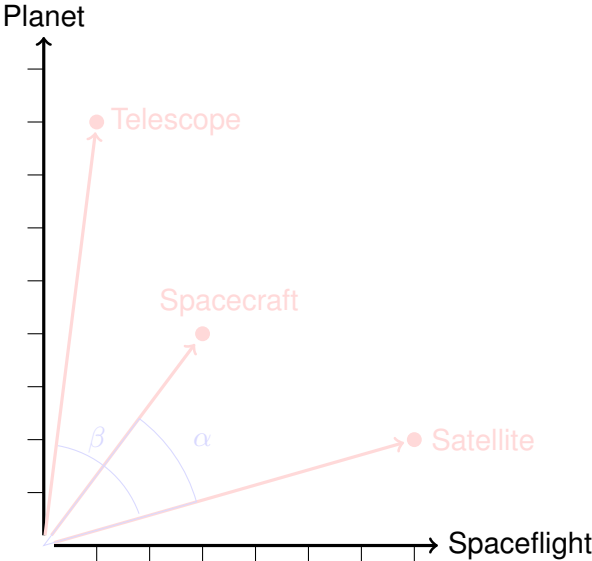
Conclusions

Feature Vectors

- ▶ For each word w we can construct a vector of feature values
- ▶ As features we use $w_1 w_2 \dots w_n$ where each w_i is a context word (closes class word in a mid frequency range).
 - ▶ We use all words with a frequency between 200 and $1 \cdot 10^6$ in our corpus
 - ▶ We have 11 080 context features.
- ▶ For a word w the value for feature w_i is the strength of the relation between w and w_i , expressed by their PPMI.
- ▶ Positive Pointwise Mutual Information (PPMI):
 $\max(0, \text{pmi}(w, w_i))$.
- ▶ Word similarity is measured by the cosine between their context vectors.



Example



Synonyms in the Dutch Parliament Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

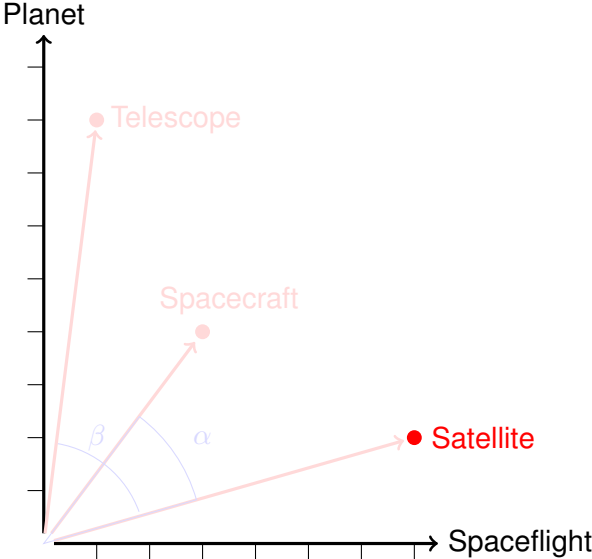
Similarity of thesaurus terms

Experiment

Conclusions



Example



Synonyms in the Dutch Parliament Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

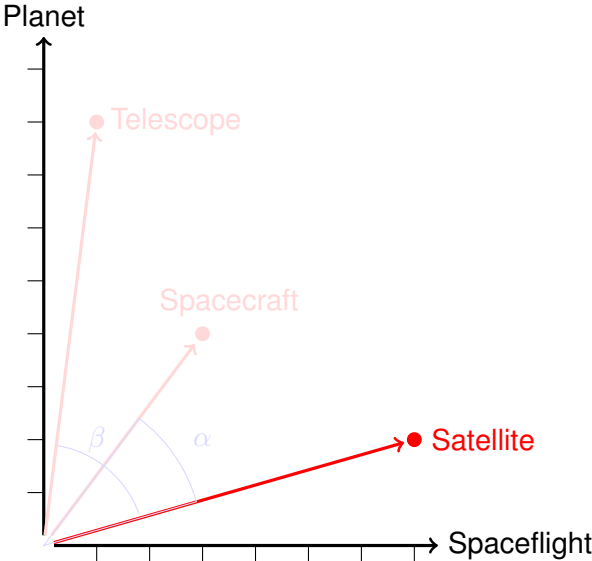
Similarity of thesaurus terms

Experiment

Conclusions



Example



Synonyms in the Dutch Parliament Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

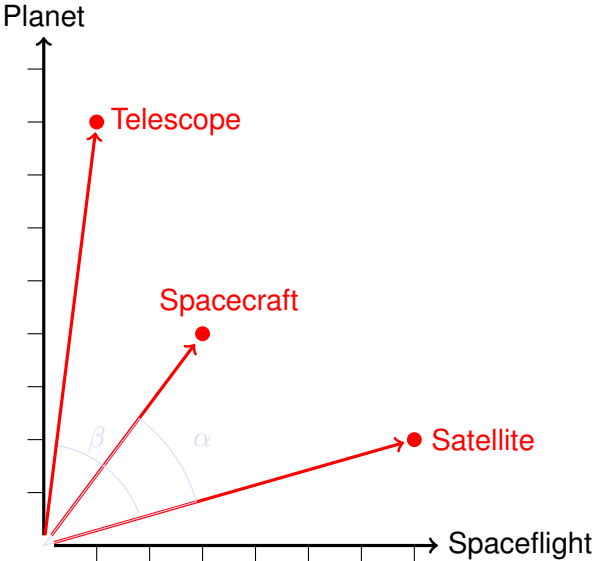
Similarity of thesaurus terms

Experiment

Conclusions



Example



Motivation

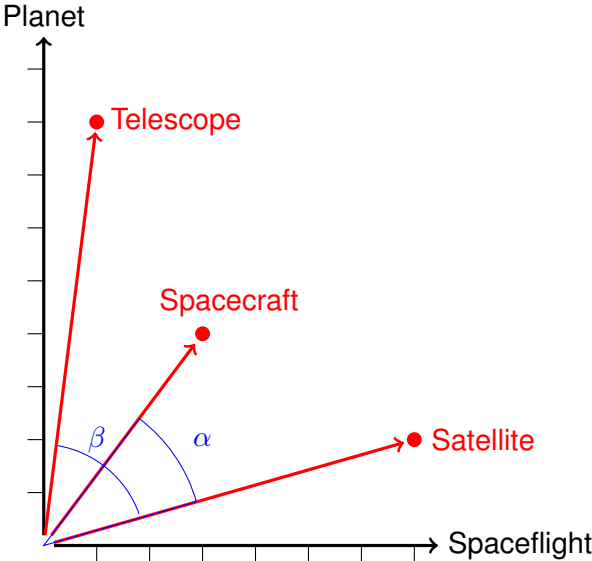
Similarity of thesaurus terms

Experiment

Conclusions



Example



Synonyms in the Dutch Parliament Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

Similarity of thesaurus terms

Experiment

Conclusions



String similarity

- ▶ Many labels are just spelling variants of the preferred label

- ▶ documentaire informatieverzorging
- ▶ documentaire informatievoorziening

- ▶ Databases
- ▶ Databanken

Trigramoverlap

- ▶ Databanken
⇒ $\{\#da, dat, ata, aba, ban, ank, nke, ken, en\}$
- ▶ Databases ⇒ $\{\#da, dat, ata, aba, bas, ase, ses, es\}$

- ▶ 13 different trigrams (union)
- ▶ 4 common trigrams (intersection)
- ▶ Trigramoverlap = $\frac{3}{14} = 0,21$

Overview

Motivation

Similarity of thesaurus terms

Experiment

Conclusions

Synonyms in the
Dutch Parliament
Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions

Experiment

- ▶ We take pairs of synonyms and non synonyms.
- ▶ Synonym: two words that are labels for the same concept.
- ▶ We compute trigram overlap and distributional similarity of word pairs using a mid-size specialized corpus.
- ▶ We use an SVM to learn the difference between synonyms and non-synonyms from those features.
- ▶ We use 10-fold cross validation for evaluation.

Word Pairs

- ▶ 3000 pairs of labels for the same concept
- ▶ 3000 pairs of labels from different concepts
 - ▶ 500 pairs of labels for related concepts
 - ▶ 500 pairs of labels for concepts related by one intermediate concept
 - ▶ 500 pairs of labels for concepts related by two intermediate concepts
 - ▶ etc.

Synonyms in the
Dutch Parliament
Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions

Word Pairs

Same Concept

- ▶ woon-werkverkeer : woonwerkverkeer (*commuting traffic*)
- ▶ vaderschapsverlof : ouderschapsverlof (*paternity/parental leave*)
- ▶ woningnood : woningzoekende (*housing shortage / house hunter*)

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions

Distance 1

- ▶ bliksemafleider : brandweer (*lightning conductor / fire brigade*)
- ▶ watersport : vaarbewijs (*aquatics / ship license*)

Distance 4

- ▶ pleziervaart : verpakkingsmateriaal (*boating / packaging material*)
- ▶ anti-raketsysteem : dienstplicht (*anti-missile system / conscription*)
- ▶ rechtswinkel : kunstverzameling (*legal aid center / arts collection*)



Corpus

- ▶ We have collected a corpus of Dutch texts from `bestanden.officielebekendmakingen.nl` from the years 2010, 2011 and 2012.
- ▶ This is the site with all official publications from the Dutch government.
- ▶ Partial overlap with the archives of the parliament.
- ▶ Due to server / connection time outs no complete years
- ▶ Raw corpus has 88,8 Million words
- ▶ We keep only unique sentences
- ▶ Resulting in a corpus of **47 Million words**.
- ▶ We lemmatized all words using the Treetagger
- ▶ We removed all stop words
- ▶ Resulting in a corpus of 40 Million words.



Features

- ▶ For each pair we compute the following features
 - ▶ Cosine between context vectors
 - ▶ Trigram overlap
- ▶ Both features are combined in a SVM

Synonyms in the
Dutch Parliament
Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions

Results I

- ▶ Using only cosine: 69% correct
- ▶ Using only trigram: 72% correct
- ▶ Using both features: 75% correct

Synonyms in the
Dutch Parliament
Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions

Results II

- ▶ If we take arbitrary word pairs and want to decide whether the terms belong to the same concept or not, the results don't carry over, because:
 - ▶ Much less than 50% of the pairs is positive
 - ▶ If we consider much more words, there are many very similar words with complete different meaning.

Overview

Motivation

Similarity of thesaurus terms

Experiment

Conclusions

Synonyms in the
Dutch Parliament
Thesaurus

Rosa Aga,
Christian Wartena,
Otto Lange,
Nelleke Aders

Motivation

Similarity of
thesaurus terms

Experiment

Conclusions

Conclusion

- ▶ DS has some potential to find new terms for a given concept
- ▶ In combination with string similarity we get quite good results
- ▶ Results could be used for proposing the most likely concept for a candidate term
- ▶ We need to improve DS!

Future Research

- ▶ Use other corpora. What is the influence of the corpus?
- ▶ Using other similarity measures. Cosine is not the best choice.
- ▶ More realistic scenarios:
 - ▶ Extract candidate terms
 - ▶ Assign candidate term to the category with the most similar labels
 - ▶ Evaluate manually



Thanks for your attention!

Questions?

The screenshot shows the Skosy (Developer Preview) application window. The title bar reads "Skosy (Developer Preview)". The menu bar includes "File", "Edit", "Preferences", and "Help". The main interface has two tabs: "Hierarchy" and "Search". The "Search" tab is active, showing a search input field with the text "vragen" and a "Search" button. Below the input field are radio buttons for "Exact search" and "Fuzzy search", and a "Search language:" dropdown menu. The search results are displayed in a list on the left side of the window, each preceded by a small Dutch flag icon. The results include:

- MONDELINGE VRAGEN**
- SCHRIFTELIJKE VRAGEN**
- vragenrecht*
- vrachtwagen*
- VERDRAGEN**
- VRACHTWAGENS**
- KAMERVragen**
- verdragenrecht*
- vrachtwagenchauffeur*
- executie verdragen*
- culturele verdragen*
- VRACHTWAGENCHAUFFEURS**
- MAATSCHAPPELIJKE VRAAGSTUKKEN**
- sociale vraagstukken*
- ADVIESRAAD INTERNATIONALE VRAAGSTUKK...**

On the right side of the window, there are two tabs: "Concept view" and "Source view". The "Concept view" tab is active, displaying the following information:

MONDELINGE VRAGEN

URI:
<http://www.tweedekamer.nl/parlementsthesaurus#2584>

MONDELINGE VRAGEN

Broader: [KAMERVragen](#)