

Challenges for Discontiguous Phrase Extraction

Dale Gerdemann

Department of Linguistics
Computational Linguistics Section
D-72074 Tübingen

Workshop on Supporting eLearning with Language
Resources and Semantic Data, 2010

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

1 Introduction

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

1 Introduction

2 Terminology

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

1 Introduction

2 Terminology

3 Toy Corpus Examples

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

1 Introduction

2 Terminology

3 Toy Corpus Examples

4 Conclusion

What's a Gappy Phrases

Recurrent patterns often occur with slight variation.

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

What's a Gappy Phrases

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Recurrent patterns often occur with slight variation.

Gappy phrase discovery is an approach for finding sequences of two phrases separated from each other by a distance not greater than d .

What's a Gappy Phrases

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Recurrent patterns often occur with slight variation.

Gappy phrase discovery is an approach for finding sequences of two phrases separated from each other by a distance not greater than d .

Examples:

- from one X to the other

What's a Gappy Phrases

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Recurrent patterns often occur with slight variation.

Gappy phrase discovery is an approach for finding sequences of two phrases separated from each other by a distance not greater than d .

Examples:

- from one X to the other
- upside [|-] down

What's a Gappy Phrases

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Recurrent patterns often occur with slight variation.

Gappy phrase discovery is an approach for finding sequences of two phrases separated from each other by a distance not greater than d .

Examples:

- from one X to the other
- upside [|-] down
- един и същи, една и съща, едно и също, едни и същи

What's a Gappy Phrases

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Recurrent patterns often occur with slight variation.

Gappy phrase discovery is an approach for finding sequences of two phrases separated from each other by a distance not greater than d .

Examples:

- from one X to the other
- upside [| -] down
- един и същи, една и съща, едно и също, едни и същи
- Bézier curve ... control point

What's a Gappy Phrases

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Recurrent patterns often occur with slight variation.

Gappy phrase discovery is an approach for finding sequences of two phrases separated from each other by a distance not greater than d .

Examples:

- from one X to the other
- upside [|-] down
- един и същи, една и съща, едно и също, едни и същи
- Bézier curve ... control point
- at the top of [her|his|his shrill little|its] voice

What's a Gappy Phrases

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Recurrent patterns often occur with slight variation.

Gappy phrase discovery is an approach for finding sequences of two phrases separated from each other by a distance not greater than d .

Examples:

- from one X to the other
- upside [| -] down
- един и същи, една и съща, едно и също, едни и същи
- Bézier curve . . . control point
- at the top of [her|his|his shrill little|its] voice
- autumn , when the leaves are [|getting] brown

How do we find such phrases?

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Algorithms for finding repeated sequences have been developed for bioinformatics/computational molecular biology. Some algorithms have been adapted to a large alphabet.

How do we find such phrases?

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Algorithms for finding repeated sequences have been developed for bioinformatics/computational molecular biology. Some algorithms have been adapted to a large alphabet.

The algorithms most interesting to NLP have been implemented in Java and are available at:
sourceforge.net/projects/saphre/.

How do we find such phrases?

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

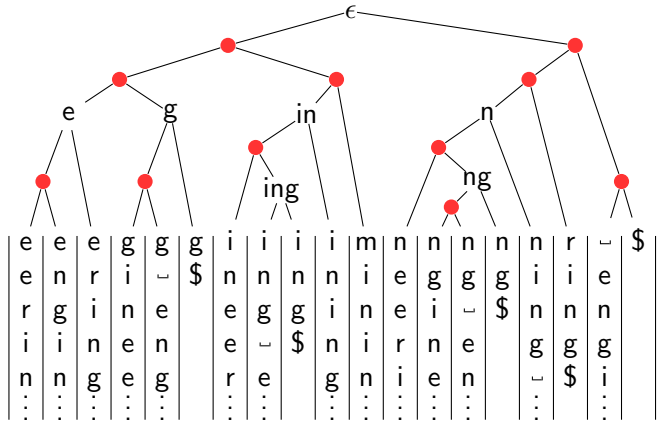
Algorithms for finding repeated sequences have been developed for bioinformatics/computational molecular biology. Some algorithms have been adapted to a large alphabet.

The algorithms most interesting to NLP have been implemented in Java and are available at:
sourceforge.net/projects/saphre/.

The programs are being actively developed, and are ideal for experimentation. Gappy phrase extraction is just one possible use.

Lcp-Interval Tree for: mining engineering

- Challenges
- Gerdemann
- Introduction
- Terminology
- Toy Corpus Examples
- Conclusion



Lcp-Interval Tree for: mining engineering

Challenges

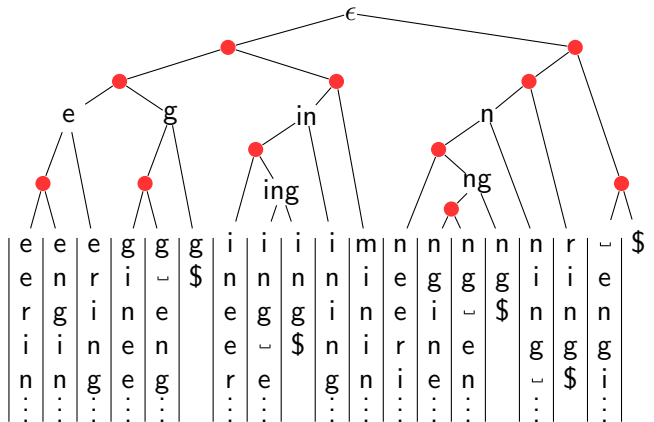
Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion



Red nodes are for binary search through possibly large alphabet; black nodes represent longest-common-prefix intervals (Kim et al., 2008).

Why is this interesting for eLearning?

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Learners can be characterized **semantically** by the concepts they express when writing: **LSA**.

Why is this interesting for eLearning?

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Learners can be characterized **semantically** by the concepts they express when writing: **LSA**.

Learners can also be characterized **phraseologically**. The phrases are indicative of Speech Genre specific to a Community of Practice. Example: Latent Semantic Indexing vs Latent Semantic Analysis.

Why is this interesting for eLearning?

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Learners can be characterized **semantically** by the concepts they express when writing: **LSA**.

Learners can also be characterized **phraseologically**. The phrases are indicative of Speech Genre specific to a Community of Practice. Example: Latent Semantic Indexing vs Latent Semantic Analysis.

Deviation from the norm indicates non-integration into the CoP: sentiment**al** analysis, gold**en** standard.

Why is this interesting for eLearning?

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Learners can be characterized **semantically** by the concepts they express when writing: **LSA**.

Learners can also be characterized **phraseologically**. The phrases are indicative of Speech Genre specific to a Community of Practice. Example: Latent Semantic Indexing vs Latent Semantic Analysis.

Deviation from the norm indicates non-integration into the CoP: sentiment**al** analysis, gold**en** standard.

Learners can be helped by a program that identifies their (perhaps subtle) linguistic peculiarities. **LTfLL project**.

Challenges

Gerdemann

Lexicography and language research

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Challenges

Gerdemann

Lexicography and language research

Introduction

Terminology

Machine translation

Toy Corpus

Examples

Conclusion

Challenges

Gerdemann

Lexicography and language research

Introduction

Terminology

Machine translation

Toy Corpus
Examples

Plagerism detection, detection of cut-and-paste problems and
detection of other intended repetitions

Conclusion

Challenges

Gerdemann

Lexicography and language research

Introduction

Terminology

Machine translation

Toy Corpus
Examples

Plagerism detection, detection of cut-and-paste problems and
detection of other intended repetitions

Conclusion

Corpus construction/consistency checking

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Lexicography and language research

Machine translation

Plagerism detection, detection of cut-and-paste problems and
detection of other intended repetitions

Corpus construction/consistency checking

Text categorization

Challenges

Gerdemann

Lexicography and language research

Introduction

Terminology

Machine translation

Toy Corpus

Examples

Plagerism detection, detection of cut-and-paste problems and detection of other intended repetitions

Conclusion

Corpus construction/consistency checking

Text categorization

Morphology induction

Challenges

Lexicography and language research

Gerdemann

Machine translation

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Plagerism detection, detection of cut-and-paste problems and detection of other intended repetitions

Corpus construction/consistency checking

Text categorization

Morphology induction

German compound noun splitting, Chinese segmentation.

A Toy Corpus

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

A corpus of 4 texts, with a sentinel at the end of each:
tortoise_the_supports_the_earth\$
put_the_cart_before_the_horse#
a_comedy_to_those_who_think%
eat_you_out_of_house_and_home&

A Toy Corpus

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

A corpus of 4 texts, with a sentinel at the end of each:

tortoise_the_supports_the_earth\$

put_the_cart_before_the_horse#

a_comedy_to_those_who_think%

eat_you_out_of_house_and_home&

A regular expression for tokenizing:

`\$\s|`

`\#\s|`

`\%\s|`

`\&\s|`

`[\w\W]`



A Toy Corpus

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

A corpus of 4 texts, with a sentinel at the end of each:
tortoise_the_supports_the_earth\$
put_the_cart_before_the_horse#
a_comedy_to_those_who_think%
eat_you_out_of_house_and_home&

A regular expression for tokenizing:

```
\$\s |  
\#\s |  
\%\s |  
\&\s |  
[\w\W]
```

Other characters, such as punctuation and paragraph boundaries, can also be treated like sentinels.

Initial Phrase Discovery

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Non-gappy phrase discovery finds the “phrase” **se**:

tortoise**se**_that_supports_the_earth\$

put_the_cart_before_the_horse**se**#

a_comedy_to_those**se**_who_think%

eat_you_out_of_house**se**_and_home&



Initial Phrase Discovery

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Non-gappy phrase discovery finds the “phrase” **se**:

tortoise_{se} that supports the earth\$

put the cart before the horse_{se}#

a comedy to those_{se} who think%

eat you out of house_{se} and home&

Note **se** is maximal (freely combinable); **se_** is supermaximal (freely combinable, and fully extended to the left and right).

leftward Search Space

Select prefix end points within distance d of the initial phrase.

to **to** **ise** _that_supports_the_earth\$
put_the_cart_before_the_ **hor** **se** #
a_comedy_to_ **tho** **se** _who_think%
eat_you_out_of_ **hou** **se** _and_home&

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

leftward Search Space

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Select prefix end points within distance d of the initial phrase.

to **tor** **se** _that_ _supports_ _the_ _earth\$
 put_ _the_ _cart_ _before_ _the_ **tor** **se** #
 a_ _comedy_ _to_ **tor** **se** _who_ _think_%
 eat_ _you_ _out_ _of_ **tor** **se** _and_ _home&

These prefixes are represented by the following numbers, and the numbers are sorted rather than the prefixes themselves.

- tor #3
- tort #4
- torto #5
- tortoi #6
- ..._the_ #57
- etc

leftward Search Space

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

We discover that **se** is preceded 3 times by **ho**, and 2 times by the longer phrase **ho**,

to**ho****se**_that_supports_the_earth\$
 put_the_cart_before_the_**ho****se**#
 a_comedy_to_**ho****se**_who_think%
 eat_you_out_of_**ho****se**_and_home&

leftward Search Space

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

We discover that **se** is preceded 3 times by **ho**, and 2 times by the longer phrase **ho**,

to**ho****se**_that_supports_the_earth\$
 put_the_cart_before_the_**ho****se**#
 a_comedy_to_**ho****se**_who_think%
 eat_you_out_of_**ho****se**_and_home&

Note: If any phrase were found from **rtoi**, **tho** and **hou**, then this phrase would have to be rejected. Such a phrase would be rediscovered when searching leftward from **se**.

Overlapping leftward Search Space

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

What happens when the leftward search spaces overlap?

tortoise_the_supports_the_earth\$

put_the_cart_before_the_horse#

a_comedy_to_those_who_think%

eat_you_out_of_house_and_home&

Overlapping leftward Search Space

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

What happens when the leftward search spaces overlap?

tortoise_the_supports_the_earth\$

put_the_cart_before_the_horse#

a_comedy_to_those_who_think%

eat_you_out_of_house_and_home&

If $d = 4$, we get the following search space:

tortoise_the_supports_the_earth\$

put_the_cart_before_the_horse#

a_comedy_to_those_who_think%

eat_you_out_of_house_and_home&

Overlapping leftward Search Space

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

What happens when the leftward search spaces overlap?

tortoise_the_supports_the_earth\$
 put_the_cart_before_the_horse#
 a_comedy_to_those_who_think%
 eat_you_out_of_house_and_home&

If $d = 4$, we get the following search space:

tortoise_the_supports_the_earth\$
 put_the_cart_before_the_horse#
 a_comedy_to_those_who_think%
 eat_you_out_of_house_and_home&

The phrase **t** is found twice within this search space. These 2 occurrences need to be matched with corresponding right parts.

Too close to the edge

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Should we “discover” that **ea** combines with **.th** to the left?

tortoise__that__supports__the__earth\$

put__the__cart__before__the__horse#

a__comedy__to__those__who__think%

ea__you__out__of__house__and__home&

Too close to the edge

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Should we “discover” that **ea** combines with **.th** to the left?

tortoise_ that_ supports_ **.th** **ea** rth\$

put_ the_ cart_ before_ the_ horse#

a_ comedy_ to_ those_ who_ **.th** think%

ea t_ you_ out_ of_ house_ and_ home&

To rule this out, we need to say that “%” and other sentinels are illegal as part of a gap.

Too close to the edge

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Should we “discover” that **ea** combines with **.th** to the left?

tortoise_ that_ supports_ **.th** **ea** rth\$

put_ the_ cart_ before_ the_ horse#

a_ comedy_ to_ those_ who_ **.th** ink%

ea t_ you_ out_ of_ house_ and_ home&

To rule this out, we need to say that “%” and other sentinels are illegal as part of a gap.

Additional illegal gap elements can be specified with a regular expression.

Alternative fillers of the gap position

Science humor collected by Joachim Verhagen

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

(asked|trying)/[(was~to explain), (who was~to)]

(making|the)/[(,~light of the), (of~use of)]

(had to|have)/[(I~also), (I~come up with)]

(matter|the problem)/[(that~cannot be), (the nature o

(don't|really don't)/[(I~care), (I~know .), (I~need t

(made|told)/[(I~it to), (were~to use)]

(obvious|possible)/[(, it is~that), (it is~. "), (som

(a specific|the)/[(do~job), (with~focus)]

(answer|one)/[(be the~to), (the correct~.)]

(building|bulb)/[(hold the~and), (on the~at the top)]

(for|to get)/[(enough~all), (enough~it to)]

(You|you)/[(~need to), (" Do~believe in), ((~could),

(first|top)/[(His~-), (of the~ten)]

(statement|story)/[(a true~.), (true~:)]

(certain|sure)/[(I can't be~.), (make~that), (never~w

Big, Little, Large, Small

See: Lynne Murphy, Semantic Relations and the Lexicon

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

(big|great)/[(a~deal), (is a~place)]

(bigger|equal)/[(are~, choose), (may be~, or)]

(big|real)/[(Coulomb got a~charge out of the), (the~deal), (you could get)]

(big|large)/[(a~ball of fire), (a~charge), (a~lump)]

(bigger|greater)/[(have a~chance of), (times~than)]

(big|good)/[(After a~meal), (a~deal), (a~f * *)]

(little|lot)/[(a~faster), (be a~more)]

(little|lot of)/[(a~fun .), (a~heat)]

(bit of|little)/[(a~fun .), (a~help from)]

(a little|the)/[(. What~acorn), (for~known), (lecture ,~old)]

(large chunk|piece)/[(a~of potassium), (such a~of)]

(a large|the)/[(at~university , the), (in~empty), (to~engineering), (to~)]

(large|real)/[(a~beaker), (a~charge)]

(large|small)/[(a~South), (a~beaker), (a~fence), (a~group of), (a~order)]

(late|small)/[(too~), (too~for it)]

(short|small)/[(a~step), (so~that you)]

(a small|the)/[(of~South), (puts~fence around)]

(problems|small children)/[(had~with him), (he had~with)]

Applications of gap alternations

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Could educational programs use gap alternations as indicators that a learner relates the concepts? Example: **machine** and **set of constraints** (PhD thesis of Jason Riggle)

Applications of gap alternations

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

Could educational programs use gap alternations as indicators that a learner relates the concepts? Example: **machine** and **set of constraints** (PhD thesis of Jason Riggle)

Could educational programs use gap alternations as indicators that a learner relates the concepts? Example: **machine** and **set of constraints** (PhD thesis of Jason Riggle)

Alternatives:

LSA Use a term-document matrix with Latent Semantic Analysis for dimensionality reduction.
Disadvantage: One needs a large number of texts from the learner.

Proximity A simple indicator of relatedness is mention of both concepts within a given radius. Ex: **Bézier curve** and **control point** in the OpenOffice user manual.

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

- Extraction of (gapped) phrases is feasible (despite the combinatorics) using suffix/prefix arrays.

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

- Extraction of (gapped) phrases is feasible (despite the combinatorics) using suffix/prefix arrays.
- Gapped and non-gapped phrases alike can be used to characterize learner, and provide formative feedback.

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

- Extraction of (gapped) phrases is feasible (despite the combinatorics) using suffix/prefix arrays.
- Gapped and non-gapped phrases alike can be used to characterize learner, and provide formative feedback.
- Gapped phrases can additionally be used to study alternative gap fillers, an indication of related concepts.

Challenges

Gerdemann

Introduction

Terminology

Toy Corpus
Examples

Conclusion

- Extraction of (gapped) phrases is feasible (despite the combinatorics) using suffix/prefix arrays.
- Gapped and non-gapped phrases alike can be used to characterize learner, and provide formative feedback.
- Gapped phrases can additionally be used to study alternative gap fillers, an indication of related concepts.
- The phrase discovery program, saphre, is freely available on Sorceforge.

Challenges

Gerdemann

Appendix

For Further
Reading

References

Dong Kyue Kim, Minhwan Kim, and Heejin Park. Linearized suffix tree: an efficient index data structure with the capabilities of suffix trees and suffix arrays. *Algorithmica*, 52 (3):350–377, 2008.