Natural Language Processing

Lecture 1
Introduction

András Simonyi, ELTE, Department of AI

2021

Course textbooks

Course textbooks

- Dan Jurafsky and James H. Martin,
  Speech and Language Processing 3rd ed.
  Draft available at
  https://web.stanford.edu/∼jurafsky/slp3.

Course textbooks

- Dan Jurafsky and James H. Martin,
  Speech and Language Processing 3rd ed.
  Draft available at
  https://web.stanford.edu/~jurafsky/slp3.
- Jacob Eisenstein,
  Natural Language Processing.
  Available at
  https://github.com/jacobeisenstein/gt-nlp-class.

Course textbooks

- Dan Jurafsky and James H. Martin,
  Speech and Language Processing 3rd ed.
  Draft available at
  https://web.stanford.edu/~jurafsky/slp3.
- Jacob Eisenstein,
  Natural Language Processing.
  Available at
  https://github.com/jacobeisenstein/gt-nlp-class.

These slides are in large part based on the first, ”Introduction” chapter of the Eisenstein book.

What is Natural Language Processing?

NLP is an interdisciplinary field concerned with making natural languages accessible to computers.

What is Natural Language Processing?

NLP is an interdisciplinary field concerned with making natural languages accessible to computers.

- Natural language in this context means the ordinary languages humans use to communicate in speech or writing, e.g., English, Chinese, Spanish etc.

What is Natural Language Processing?

NLP is an interdisciplinary field concerned with making natural languages accessible to computers.

- Natural language in this context means the ordinary languages humans use to communicate in speech or writing, e.g., English, Chinese, Spanish etc.
- Being able to access natural language covers a wide range of capabilities, with important areas being

What is Natural Language Processing?

NLP is an interdisciplinary field concerned with making natural languages accessible to computers.

- Natural language in this context means the ordinary languages humans use to communicate in speech or writing, e.g., English, Chinese, Spanish etc.
- Being able to access natural language covers a wide range of capabilities, with important areas being
  - communication: the ability to accept input and produce output in natural language;

What is Natural Language Processing?

NLP is an interdisciplinary field concerned with making natural languages accessible to computers.

- Natural language in this context means the ordinary languages humans use to communicate in speech or writing, e.g., English, Chinese, Spanish etc.
- Being able to access natural language covers a wide range of capabilities, with important areas being
  - communication: the ability to accept input and produce output in natural language;
  - understanding: being able to access and utilize the informational and emotional content;

What is Natural Language Processing?

NLP is an interdisciplinary field concerned with making natural languages accessible to computers.

- Natural language in this context means the ordinary languages humans use to communicate in speech or writing, e.g., English, Chinese, Spanish etc.
- Being able to access natural language covers a wide range of capabilities, with important areas being
  - communication: the ability to accept input and produce output in natural language;
  - understanding: being able to access and utilize the informational and emotional content;
  - linguistic assistance: the ability to help humans to express themselves linguistically.

Related fields

Computational linguistics

The scientific study of language using computational methods.

Computational linguistics

The scientific study of language using computational methods.

- Perhaps the closest field to NLP, but with a different focus: NLP is *not* concerned with theoretical insights into natural languages *per se* but only with the design and analysis of methods useful for computational language processing.

Computational linguistics

The scientific study of language using computational methods.

- Perhaps the closest field to NLP, but with a different focus: NLP is *not* concerned with theoretical insights into natural languages *per se* but only with the design and analysis of methods useful for computational language processing.
- Instead of direct implementations of theoretical ideas, it often provides *architectural inspiration* for NLP systems.

Artificial intelligence (AI)

There is obviously a large overlap between the NLP objective and AI's goal of building intelligent systems:

Artificial intelligence (AI)

There is obviously a large overlap between the NLP objective and AI's goal of building intelligent systems:

- language use is strongly interdependent with the conceptual, representational and reasoning capabilities required for being intelligent,

Artificial intelligence (AI)

There is obviously a large overlap between the NLP objective and AI's goal of building intelligent systems:

- language use is strongly interdependent with the conceptual, representational and reasoning capabilities required for being intelligent,
- in practice, large-scale knowledge acquisition is also impossible without the ability to extract information from natural language input.

Artificial intelligence (AI)

There is obviously a large overlap between the NLP objective and AI's goal of building intelligent systems:

- language use is strongly interdependent with the conceptual, representational and reasoning capabilities required for being intelligent,
- in practice, large-scale knowledge acquisition is also impossible without the ability to extract information from natural language input.

The above characteristics make especially the AI subfields Knowledge Representation and Reasoning very relevant for NLP.

Machine learning

Modern NLP relies on machine learning techniques to a huge extent, in fact, in recent years the linguistic applications of general ML methods have dominated the area.

Machine learning

Modern NLP relies on machine learning techniques to a huge extent, in fact, in recent years the linguistic applications of general ML methods have dominated the area.

- Mostly supervised or semi-supervised methods are utilized, but the use of reinforcement learning is also increasing.

Machine learning

Modern NLP relies on machine learning techniques to a huge extent, in fact, in recent years the linguistic applications of general ML methods have dominated the area.

- Mostly supervised or semi-supervised methods are utilized, but the use of reinforcement learning is also increasing.
- Texts are sequences of discrete symbols, so ML models capable of dealing with this type of input (and output in case of generation) are needed.

Speech processing

The processing and generation of acoustic speech signals is traditionally not considered part of NLP, which is concerned primarily with *texts*, but is obviusly closely related:

Speech processing

The processing and generation of acoustic speech signals is traditionally not considered part of NLP, which is concerned primarily with *texts*, but is obviusly closely related:

- Speech2text provides input for NLP applications,

Speech processing

The processing and generation of acoustic speech signals is traditionally not considered part of NLP, which is concerned primarily with *texts*, but is obviusly closely related:

- Speech2text provides input for NLP applications,
- NLP apps provide input for speech synthesis;

Speech processing

The processing and generation of acoustic speech signals is traditionally not considered part of NLP, which is concerned primarily with *texts*, but is obviusly closely related:

- Speech2text provides input for NLP applications,
- NLP apps provide input for speech synthesis;
- Both processing and synthesizing speech requires linguistic knowledge that is also relevant for NLP: especially *language modeling* has a central role in both areas.

Applications

Application examples

- machine translation,

Application examples

- machine translation,
- document retrieval: retrieving free-text documents matching a user query,

Application examples

- machine translation,
- document retrieval: retrieving free-text documents matching a user query,
- question answering, e.g., smart phone assistants' ability to answer questions,

Application examples

- machine translation,
- document retrieval: retrieving free-text documents matching a user query,
- question answering, e.g., smart phone assistants' ability to answer questions,
- text classification, e.g. detecting e-mail spam,

Application examples

- machine translation,
- document retrieval: retrieving free-text documents matching a user query,
- question answering, e.g., smart phone assistants' ability to answer questions,
- text classification, e.g. detecting e-mail spam,
- chatbots, e.g., a chatbot for buying a train ticket,

Application examples

- machine translation,
- document retrieval: retrieving free-text documents matching a user query,
- question answering, e.g., smart phone assistants' ability to answer questions,
- text classification, e.g. detecting e-mail spam,
- chatbots, e.g., a chatbot for buying a train ticket,
- spell checking and grammar checking,

Application examples

- machine translation,
- document retrieval: retrieving free-text documents matching a user query,
- question answering, e.g., smart phone assistants' ability to answer questions,
- text classification, e.g. detecting e-mail spam,
- chatbots, e.g., a chatbot for buying a train ticket,
- spell checking and grammar checking,
- auto-completion for free-text input,

Application examples

- machine translation,
- document retrieval: retrieving free-text documents matching a user query,
- question answering, e.g., smart phone assistants' ability to answer questions,
- text classification, e.g. detecting e-mail spam,
- chatbots, e.g., a chatbot for buying a train ticket,
- spell checking and grammar checking,
- auto-completion for free-text input,
- document summarization,

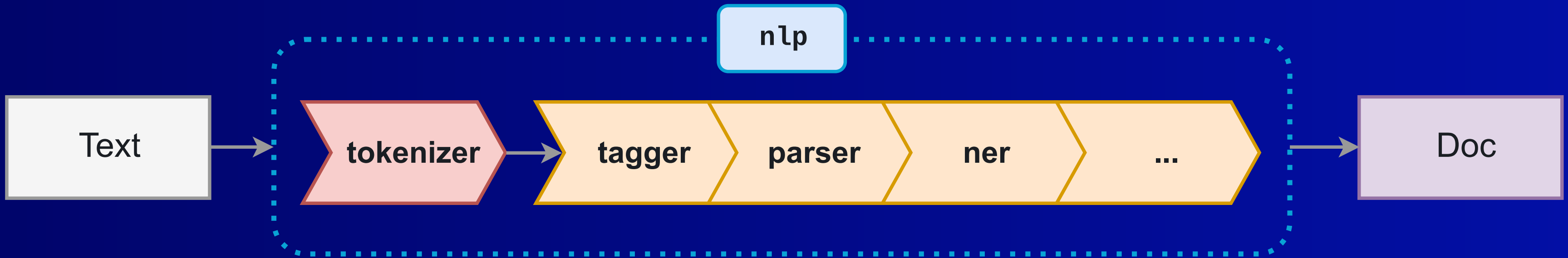
Application examples

- machine translation,
- document retrieval: retrieving free-text documents matching a user query,
- question answering, e.g., smart phone assistants' ability to answer questions,
- text classification, e.g. detecting e-mail spam,
- chatbots, e.g., a chatbot for buying a train ticket,
- spell checking and grammar checking,
- auto-completion for free-text input,
- document summarization,
- text generation from structured data (from stock exchange news to error messages).

Central themes

Pipeline vs end-to-end architectures

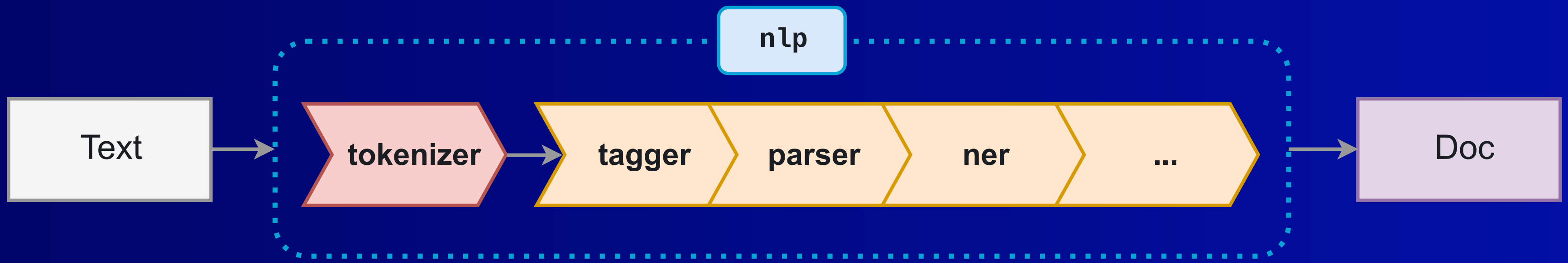
An influential view of NLP considers its core task to provide a pipeline of modules that successively produce general-purpose linguistic analyses, each module building on the outputs of the previous ones:



(Figure from the documentation of the spaCy NLP library.)

Pipeline vs end-to-end architectures

An influential view of NLP considers its core task to provide a pipeline of modules that successively produce general-purpose linguistic analyses, each module building on the outputs of the previous ones:



(Figure from the documentation of the spaCy NLP library.)

Specialized NLP applications are then built as relatively simple additions on top of elements of this universal pipeline.

Pipeline vs end-to-end cont.

The opposite view concentrates on building NLP applications as end-to-end machine learning models that learn to transform the raw input to the required output without specialized linguistic analyzer modules.

Pipeline vs end-to-end cont.

The opposite view concentrates on building NLP applications as end-to-end machine learning models that learn to transform the raw input to the required output without specialized linguistic analyzer modules.

State-of-the art NLP applications frequently fall in between these two extremes: they use some universal analyzer modules, e.g., for word segmentation or stemming, and also rely on ML models that skip some of the traditional pipeline steps to produce the required output.

Transfer learning

An interesting, relatively recent, development is the appearance of neural models that are end-to-end pretrained on unsupervised tasks on very large text collections, and can act as a replacement for the traditional processing pipelines:

Transfer learning

An interesting, relatively recent, development is the appearance of neural models that are end-to-end pretrained on unsupervised tasks on very large text collections, and can act as a replacement for the traditional processing pipelines:

- Specialized models can be built by adding a few very shallow layers to the architecture but keeping the pretrained weights perhaps with a bit of fine-tuning.

Transfer learning

An interesting, relatively recent, development is the appearance of neural models that are end-to-end pretrained on unsupervised tasks on very large text collections, and can act as a replacement for the traditional processing pipelines:

- Specialized models can be built by adding a few very shallow layers to the architecture but keeping the pretrained weights perhaps with a bit of fine-tuning.
- It seems that some components of traditional pipeline have neural analogues in these models: certain layers seem to learn (more) morphology, others semantics etc.

Transfer learning

An interesting, relatively recent, development is the appearance of neural models that are end-to-end pretrained on unsupervised tasks on very large text collections, and can act as a replacement for the traditional processing pipelines:

- Specialized models can be built by adding a few very shallow layers to the architecture but keeping the pretrained weights perhaps with a bit of fine-tuning.
- It seems that some components of traditional pipeline have neural analogues in these models: certain layers seem to learn (more) morphology, others semantics etc.

Learning and search

A large number of supervised NLP tasks which we will encounter can be formulated as an optimization problem of the form

$$\hat{y} = \operatorname*{argmax}_{y \in Y(x)} \Psi_\theta(x, y)$$

Learning and search

A large number of supervised NLP tasks which we will encounter can be formulated as an optimization problem of the form

$$\hat{y} = \operatorname*{argmax}_{y \in Y(x)} \Psi_\theta(x, y)$$

where

- $x \in X$ and $Y(x)$ are the task's input and potential outputs,

Learning and search

A large number of supervised NLP tasks which we will encounter can be formulated as an optimization problem of the form

$$\hat{y} = \operatorname*{argmax}_{y \in Y(x)} \Psi_\theta(x, y)$$

where

- $x \in X$ and $Y(x)$ are the task's input and potential outputs,
- $\Psi_\theta : X \times Y \to \mathbb{R}$ is a scoring function or model that assigns scores to $\langle x, y \rangle$ input-output pairs and is parametrized by a θ vector,

Learning and search

A large number of supervised NLP tasks which we will encounter can be formulated as an optimization problem of the form

$$\hat{y} = \operatorname*{argmax}_{y \in Y(x)} \Psi_\theta(x, y)$$

where

- $x \in X$ and $Y(x)$ are the task's input and potential outputs,
- $\Psi_\theta : X \times Y \to \mathbb{R}$ is a scoring function or model that assigns scores to $\langle x, y \rangle$ input-output pairs and is parametrized by a θ vector, and
- $\hat{y}$ is the predicted output.

Learning and search cont.

For instance,

- X could contain movie reviews and Y the sentiment labels Positive, Negative and Neutral, and Ψ_θ could be a function assigning probabilities to the possible sentiment labelings of the reviews.

Learning and search cont.

For instance,

- X could contain movie reviews and Y the sentiment labels POSITIVE, NEGATIVE and NEUTRAL, and Ψ_θ could be a function assigning probabilities to the possible sentiment labelings of the reviews.
- Also, X could be the set of German texts and Y their potential English translations, with Ψ_θ assigning translation quality scores to the candidates.

Learning and search cont.

For instance,

- X could contain movie reviews and Y the sentiment labels POSITIVE, NEGATIVE and NEUTRAL, and Ψ_θ could be a function assigning probabilities to the possible sentiment labelings of the reviews.
- Also, X could be the set of German texts and Y their potential English translations, with Ψ_θ assigning translation quality scores to the candidates.

Learning and search cont.

This formulation makes it possible to factorize the problem into two optimization subproblems solved by two distinct modules:

Learning and search cont.

This formulation makes it possible to factorize the problem into two optimization subproblems solved by two distinct modules:

- Learning: Finding the optimal θ parameters. This is typically done by optimizing θ on a large supervised data set $\{\langle x_i, y_i \rangle\}_{i=1}^{N}$ using numerical optimization methods.

Learning and search cont.

This formulation makes it possible to factorize the problem into two optimization subproblems solved by two distinct modules:

- Learning: Finding the optimal θ parameters. This is typically done by optimizing θ on a large supervised data set $\{\langle x_i, y_i \rangle\}_{i=1}^{N}$ using numerical optimization methods.
- Search: Finding the best scoring y for a specific x, i.e., computing the value of the argmax in the formula. Since the search space $Y(x)$ is often large because the potential ys have a complex structure (think, e.g., of a parse tree), this problem frequently requires combinatorial optimization.

Semantic perspectives: relational

Consider the utterance

My uncle's bought a cat. He's perhaps the most obnoxious animal I've ever met.

Semantic perspectives: relational

Consider the utterance

My uncle's bought a cat. He's perhaps the most obnoxious animal I've ever met.

How do we know that "animal" is said about the mentioned cat? One factor is that we know that *cat* is a subcategory of *animal*: they are connected by the IS_A relationship.

Semantic perspectives: relational

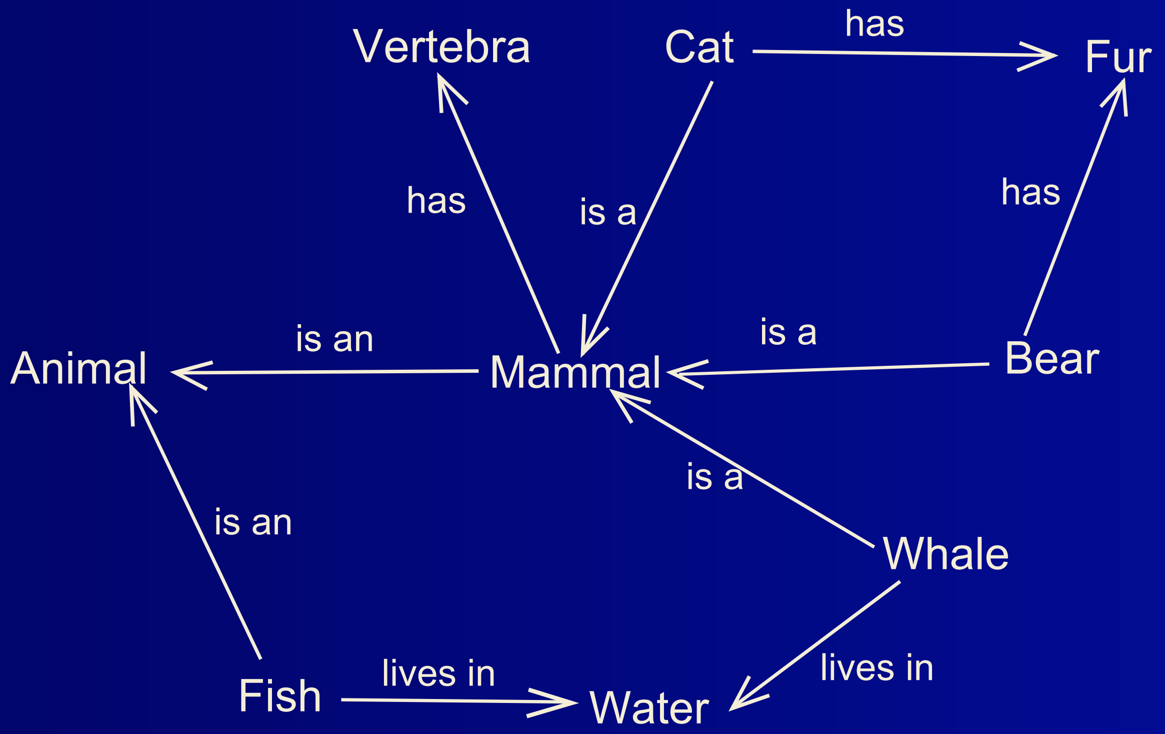
Consider the utterance

My uncle's bought a cat. He's perhaps the most obnoxious animal I've ever met.

How do we know that "animal" is said about the mentioned cat? One factor is that we know that *cat* is a subcategory of *animal*: they are connected by the IS_A relationship.

The relational perspective concentrates on these semantic/conceptual links between the senses of expressions, which together constitute semantic networks:

Semantic perspectives: relational



(A semantic network fragment. Figure from Wikipedia: Semantic Networks.)

Lexical semantic ontologies like WordNet and FrameNet are attempts to enumerate the semantic relations between a large number of word senses.

Semantic perspectives: compositionality

The relational view sees meanings as atomic nodes in a network. The compositional perspective, in contrast, analyzes an expression's meaning according to its internal composition.

Semantic perspectives: compositionality

The relational view sees meanings as atomic nodes in a network. The compositional perspective, in contrast, analyzes an expression's meaning according to its internal composition.

E.g., the decomposition

un|*bear*|*able*|*s*

Semantic perspectives: compositionality

The relational view sees meanings as atomic nodes in a network. The compositional perspective, in contrast, analyzes an expression's meaning according to its internal composition.

E.g., the decomposition

un|*bear*|*able*|*s*

allows us to see the meaning of *unbearables* as being composed of the meanings of its parts *un*, *bear*, *able* and *s*.

Semantic perspectives: compositionality

The principle of compositionality:

Semantic perspectives: compositionality

The principle of compositionality:

The meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them. [1]

Semantic perspectives: compositionality

The principle of compositionality:

The meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them. [1]

The principle can be applied to larger linguistic units than words: sentences or even paragraphs etc. One (traditional) approach is to represent meanings with logical formulas and associate syntactic syntactic rules of combination with semantic/logical ones:

[1] Wikipedia: Principle of Compositionality.

Semantic perspectives: compositionality

John visits Julie (S)

John (NP) *visits Julie* (VP)

visits (VT) *Julie* (NP)

Semantic perspectives: compositionality

John visits Julie (S)

John (NP) *visits Julie* (VP)

visits (VT) *Julie* (NP)

VISITS(JOHN,JULIE)

JOHN $\lambda x.$VISITS$(x,$JULIE$)$

$\lambda y.\lambda x.$VISITS(x, y) JULIE

Semantic perspectives: distributional

What does “bardiwac” mean?[2]

Semantic perspectives: distributional

What does “bardiwac” mean?[2]

- He handed her a glass of bardiwac.

Semantic perspectives: distributional

What does “bardiwac” mean?[2]

- He handed her a glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.

Semantic perspectives: distributional

What does “bardiwac” mean?[2]

- He handed her a glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.
- The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.

Semantic perspectives: distributional

What does “bardiwac” mean?[2]

- He handed her a glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.
- The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.
- Nigel’s face flushed from too much bardiwac.

Semantic perspectives: distributional

What does “bardiwac” mean?[2]

- He handed her a glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.
- The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.
- Nigel’s face flushed from too much bardiwac.
- Malbec is one of the lesser-known bardiwac grapes.

Semantic perspectives: distributional

What does “bardiwac” mean?[2]

- He handed her a glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.
- The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.
- Nigel’s face flushed from too much bardiwac.
- Malbec is one of the lesser-known bardiwac grapes.
- I dined off bread, cheese and this excellent bardiwac.

Semantic perspectives: distributional

What does “bardiwac” mean?[2]

- He handed her a glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.
- The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.
- Nigel’s face flushed from too much bardiwac.
- Malbec is one of the lesser-known bardiwac grapes.
- I dined off bread, cheese and this excellent bardiwac.

$\Rightarrow$ Bardiwac is a heavy red alcoholic beverage made from grapes.

[2] The example is from Stefan Evert’s Distributional semantics slides.

Semantic perspectives: distributional

Even if we don't know the place of "bardiwac" in a semantic network nor the meanings of its parts, the *contexts* in which it occurs provide a large amount of information about it's meaning.

Semantic perspectives: distributional

Even if we don’t know the place of “bardiwac” in a semantic network nor the meanings of its parts, the *contexts* in which it occurs provide a large amount of information about it’s meaning.

The distributional hypothesis:

Semantic perspectives: distributional

Even if we don’t know the place of “bardiwac” in a semantic network nor the meanings of its parts, the *contexts* in which it occurs provide a large amount of information about it’s meaning.

The distributional hypothesis:

- “You shall know a word by the company it keeps.” [3]

Semantic perspectives: distributional

Even if we don't know the place of "bardiwac" in a semantic network nor the meanings of its parts, the *contexts* in which it occurs provide a large amount of information about it's meaning.

The distributional hypothesis:

- "You shall know a word by the company it keeps." [3]
- "Linguistic items with similar distributions have similar meanings." [4]

[3] J.R. Firth, *Papers in Linguistics 1934–1951 (1957).*

[4] Wikipedia: Distributional semantics.

Semantic perspectives: distributional

An important practical advantage of the distributional approach to meaning is that it makes it possible to learn the semantics of words automatically from large but unlabeled text collections, no expert knowledge and annotations are needed.

Semantic perspectives: distributional

An important practical advantage of the distributional approach to meaning is that it makes it possible to learn the semantics of words automatically from large but unlabeled text collections, no expert knowledge and annotations are needed.

The approach is not without limitations, of course:

Semantic perspectives: distributional

An important practical advantage of the distributional approach to meaning is that it makes it possible to learn the semantics of words automatically from large but unlabeled text collections, no expert knowledge and annotations are needed.

The approach is not without limitations, of course:

- has problems with rare words; and

Semantic perspectives: distributional

An important practical advantage of the distributional approach to meaning is that it makes it possible to learn the semantics of words automatically from large but unlabeled text collections, no expert knowledge and annotations are needed.

The approach is not without limitations, of course:

- has problems with rare words; and
- learns the similarities without providing any explanation *why* these distributions are similar.