KI-Campus
Die Lernplattform
für Künstliche Intelligenz

Deutsches
Forschungszentrum
für Künstliche
Intelligenz GmbH

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Keyphrase Extraction

Salar Mohtaj | DFKI

Keyphrase extraction

- What is keyphrase extraction
- Why is keyphrase extraction important
- Classical keyphrase extraction methods
- Neural keyphrase extraction
- Evaluation of automatic keyphrase extraction

Keyphrase extraction

- What is keyphrase extraction
- Why is keyphrase extraction important
- Classical keyphrase extraction methods
- Neural keyphrase extraction
- Evaluation of automatic keyphrase extraction

What is keyphrase extraction

- Keyphrase extraction is the automated process of extracting the most relevant words/phrases and expressions from text
- Automatic keyphrase extraction (AKE) is the task to identify a small set of words, key phrases, keywords, or key segments from a document that can describe the meaning of the document

What is keyphrase extraction

- Due to the exponential growth of textual data and web sources, an automatic mechanism required to identify relevant information embedded within them
- It helps summarize the content of texts and recognize the main topics discussed

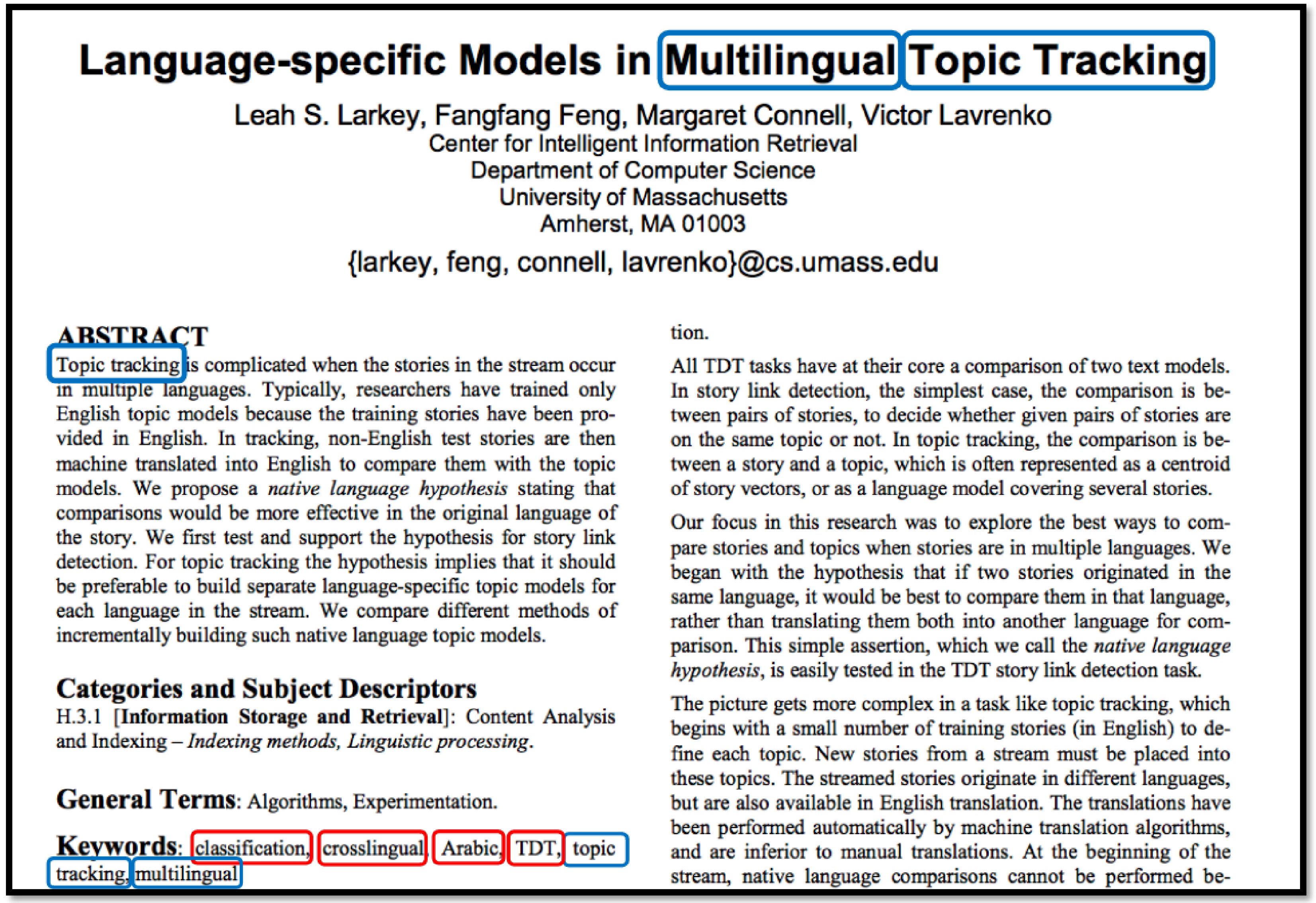
What is keyphrase extraction

- A **keyword** is a single word that represent the main topic of the text.
  A **keyphrase** is a sequence of one or more words that are considered highly relevant

https://monkeylearn.com

Why is keyphrase extraction difficult

- Some documents cover different topics
- Keyphrases are not necessarily the most frequent phrases
- Sometimes the Keyphrases don't present in the document



Language-specific Models in Multilingual Topic Tracking

Leah S. Larkey, Fangfang Feng, Margaret Connell, Victor Lavrenko
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{larkey, feng, connell, lavrenko}@cs.umass.edu

ABSTRACT

Topic tracking is complicated when the stories in the stream occur in multiple languages. Typically, researchers have trained only English topic models because the training stories have been provided in English. In tracking, non-English test stories are then machine translated into English to compare them with the topic models. We propose a *native language hypothesis* stating that comparisons would be more effective in the original language of the story. We first test and support the hypothesis for story link detection. For topic tracking the hypothesis implies that it should be preferable to build separate language-specific topic models for each language in the stream. We compare different methods of incrementally building such native language topic models.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Indexing methods, Linguistic processing*.

General Terms: Algorithms, Experimentation.

Keywords: classification, crosslingual, Arabic, TDT, topic tracking, multilingual

tion.

All TDT tasks have at their core a comparison of two text models. In story link detection, the simplest case, the comparison is between pairs of stories, to decide whether given pairs of stories are on the same topic or not. In topic tracking, the comparison is between a story and a topic, which is often represented as a centroid of story vectors, or as a language model covering several stories.

Our focus in this research was to explore the best ways to compare stories and topics when stories are in multiple languages. We began with the hypothesis that if two stories originated in the same language, it would be best to compare them in that language, rather than translating them both into another language for comparison. This simple assertion, which we call the *native language hypothesis*, is easily tested in the TDT story link detection task.

The picture gets more complex in a task like topic tracking, which begins with a small number of training stories (in English) to define each topic. New stories from a stream must be placed into these topics. The streamed stories originate in different languages, but are also available in English translation. The translations have been performed automatically by machine translation algorithms, and are inferior to manual translations. At the beginning of the stream, native language comparisons cannot be performed be-

Meng, Rui, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. "Deep keyphrase generation." arXiv preprint arXiv:1704.06879 (2017).

Keyphrase extraction

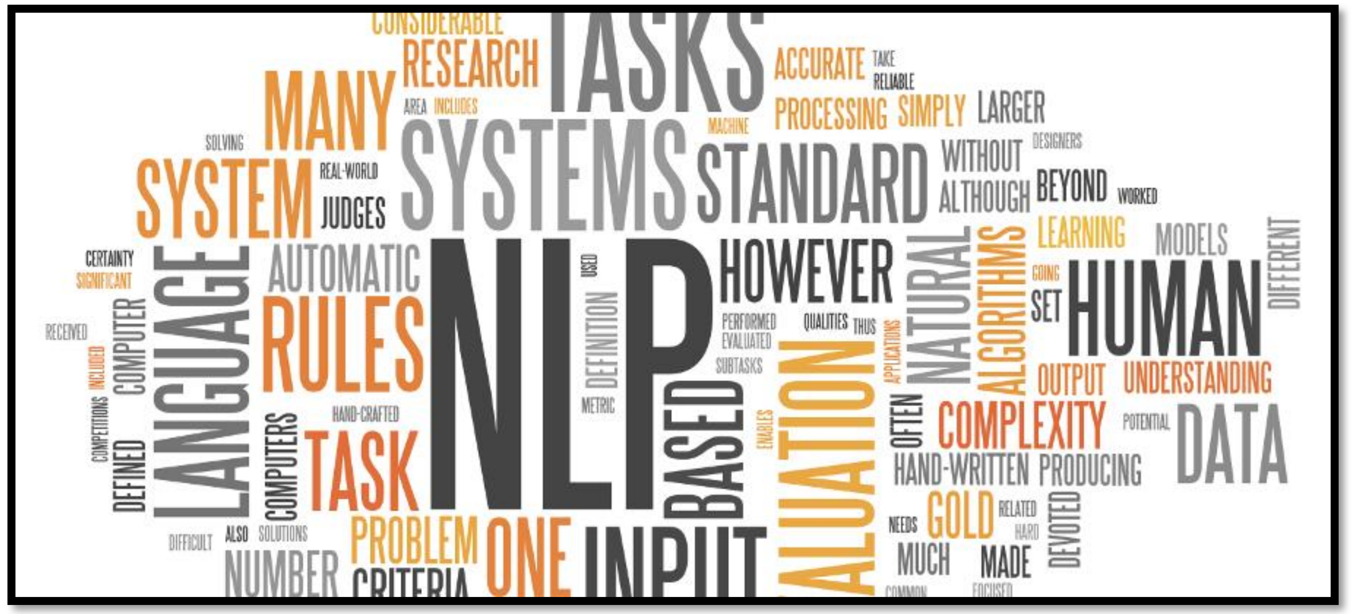
- What is keyphrase extraction
- Why is keyphrase extraction important
- Classical keyphrase extraction methods
- Neural keyphrase extraction
- Evaluation of automatic keyphrase extraction

Why is keyphrase extraction important

- Keyphrases in a document provide important information about the content of the document
  - They can help users search through information more efficiently or ***decide*** whether to read a document
- Considering that most of the data we generate every day is unstructured, businesses need automated keyphrase extraction to help them ***process*** and ***analyze*** customer data in a more efficient manner
- Keyphrase extraction can be considered as the ***core technology*** of most of the text processing applications

Why is keyphrase extraction important

- Many NLP applications can take advantage of key words/phrases
  - Automatic summarization
  - Text classification
  - Text clustering
  - Automatic filtering
  - Topic detection and tracking
  - Information visualization



http://erikburger.nl

Keyphrase extraction

- What is keyphrase extraction
- Why is keyphrase extraction important
- Classical keyphrase extraction methods
- Neural keyphrase extraction
- Evaluation of automatic keyphrase extraction

Classical keyphrase extraction methods

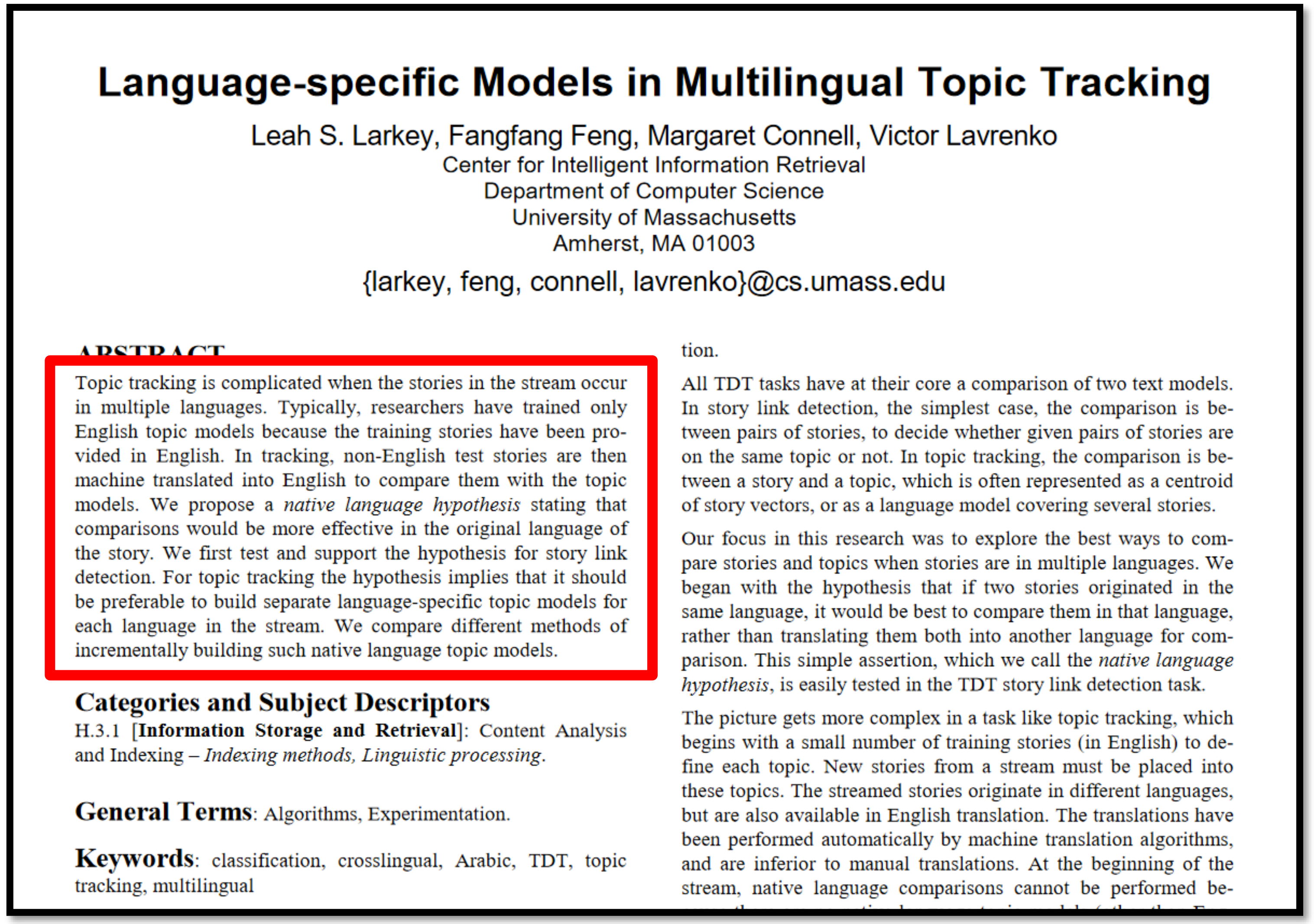
- Generally, classical systems identify a set of words and phrases called ***candidates*** that could convey the topical content of a document
- Then these candidates are ***scored*** and ***ranked***
- Finally, the ***best*** ones are selected as a document’s ***keyphrases***

Classical keyphrase extraction methods

- Candidate identification
- Keyphrase selection
  - Unsupervised approaches
  - Supervised models

Candidate identification

- Selecting candidate words and phrases
- Using ***heuristic rules*** to extract a set of phrases and words as candidate keyphrases
- The idea is to keep the number of candidates to ***a minimum***
- Still keeping ***high recall*** and don't miss good candidates



Language-specific Models in Multilingual Topic Tracking

Leah S. Larkey, Fangfang Feng, Margaret Connell, Victor Lavrenko
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{larkey, feng, connell, lavrenko}@cs.umass.edu

ABSTRACT

Topic tracking is complicated when the stories in the stream occur in multiple languages. Typically, researchers have trained only English topic models because the training stories have been provided in English. In tracking, non-English test stories are then machine translated into English to compare them with the topic models. We propose a *native language hypothesis* stating that comparisons would be more effective in the original language of the story. We first test and support the hypothesis for story link detection. For topic tracking the hypothesis implies that it should be preferable to build separate language-specific topic models for each language in the stream. We compare different methods of incrementally building such native language topic models.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Indexing methods, Linguistic processing.*

General Terms: Algorithms, Experimentation.

Keywords: classification, crosslingual, Arabic, TDT, topic tracking, multilingual

tion.

All TDT tasks have at their core a comparison of two text models. In story link detection, the simplest case, the comparison is between pairs of stories, to decide whether given pairs of stories are on the same topic or not. In topic tracking, the comparison is between a story and a topic, which is often represented as a centroid of story vectors, or as a language model covering several stories.

Our focus in this research was to explore the best ways to compare stories and topics when stories are in multiple languages. We began with the hypothesis that if two stories originated in the same language, it would be best to compare them in that language, rather than translating them both into another language for comparison. This simple assertion, which we call the *native language hypothesis*, is easily tested in the TDT story link detection task.

The picture gets more complex in a task like topic tracking, which begins with a small number of training stories (in English) to define each topic. New stories from a stream must be placed into these topics. The streamed stories originate in different languages, but are also available in English translation. The translations have been performed automatically by machine translation algorithms, and are inferior to manual translations. At the beginning of the stream, native language comparisons cannot be performed be-

Candidate identification

- Typical heuristics
  - Removing stop words
  - Allowing words with certain part-of-speech tags (e.g., nouns, adjectives, verbs)
  - Using external knowledge bases like WordNet or Wikipedia as a reference source of keyphrases
    - Phrases which appear in Wikipedia article titles
  - Generating n-grams for different ranges of N
    - Extracting noun phrases based on grammatical rules

The **election-year politics** are annoying for **many people**.

Candidate Identification

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

Candidate Identification

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

Keyphrase selection

- In this step the idea is to select good candidates among the whole list of candidates
- A very simple approach could be weighing candidates based on frequency statistics like TF-IDF
  - Best keyphrases are not necessarily the most frequent within a document
- Two different approaches
  - Unsupervised approaches
  - Supervised models

Unsupervised keyphrase selection

- The idea is to select the best keyphrases from the candidate list without relying on labeled data (training data)
  - Graph-based ranking method
  - Topic-based clustering

Hasan, Kazi Saidul, and Vincent Ng. "Automatic keyphrase extraction: A survey of the state of the art." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1262-1273. 2014.

Graph-based ranking method

- An important candidate is related to:
  1. A large number of other candidates
  2. Candidates which are important

- A document is represented as a graph
  - Nodes are candidate keyphrases
  - Edges connect related candidates

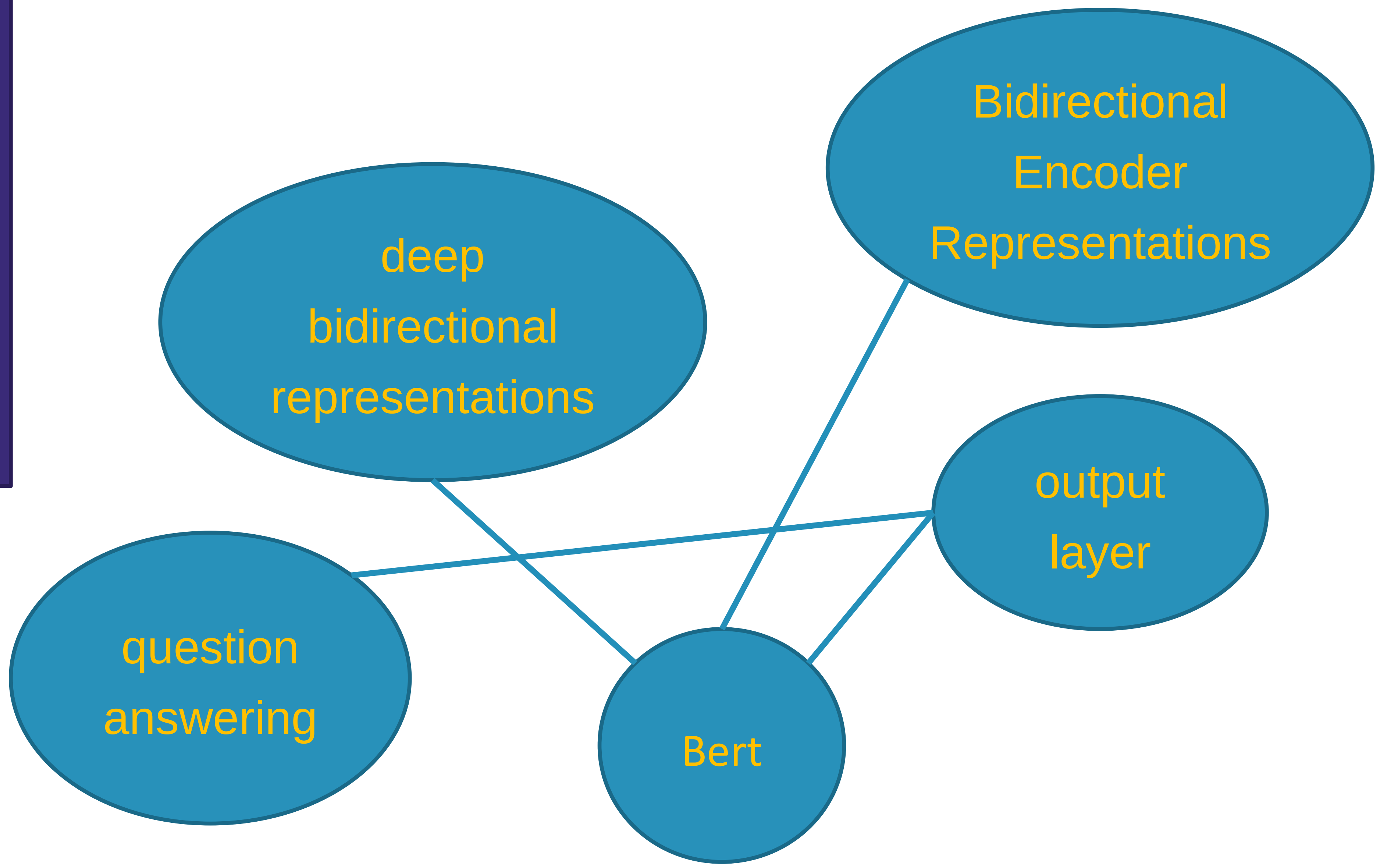
Hasan, Kazi Saidul, and Vincent Ng. "Automatic keyphrase extraction: A survey of the state of the art." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1262-1273. 2014.

Graph-based ranking method

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

Graph-based ranking method

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.



Graph-based ranking method

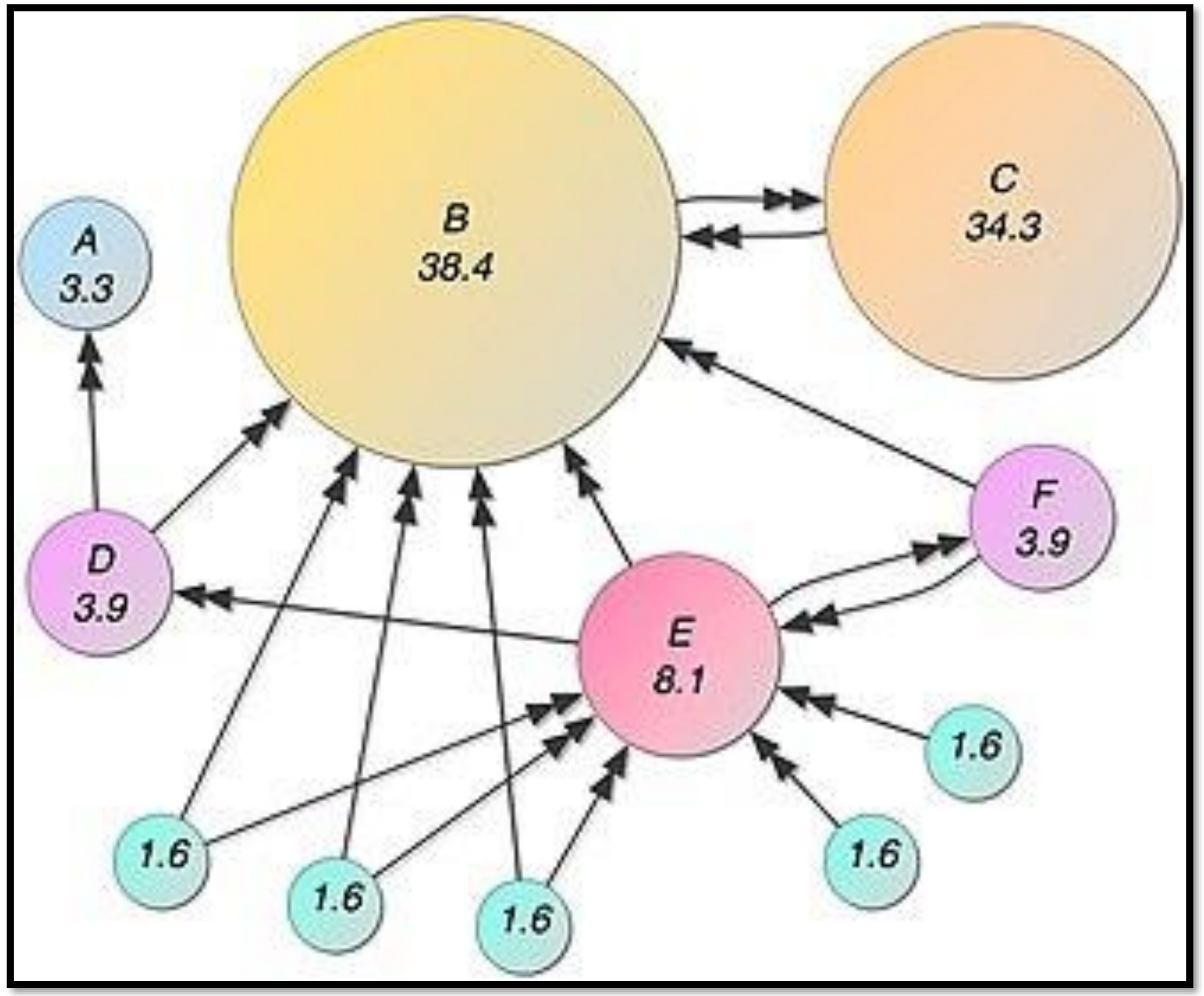
- An important candidate is related to:
1. A large number of other candidates
2. Candidates which are important

- A document is represented as a graph
  - Nodes are candidate keyphrases
  - Edges connect related candidates
  - Then, a graph-based ranking algorithm, such as PageRank, is run over the graph
    - The highest-scoring terms are keyphrases

Hasan, Kazi Saidul, and Vincent Ng. "Automatic keyphrase extraction: A survey of the state of the art." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1262-1273. 2014.

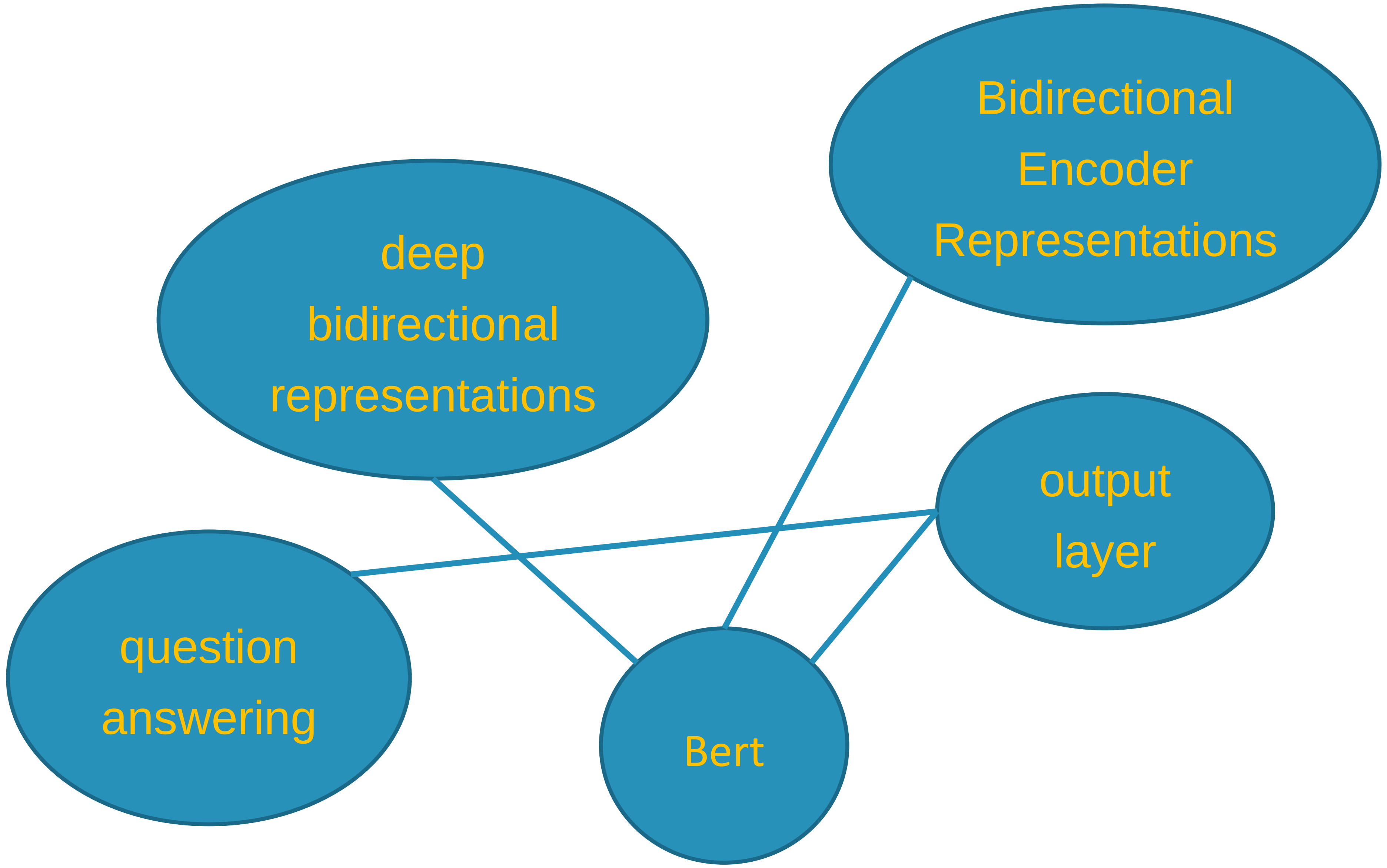
Graph-based ranking method

- PageRank
  - PageRank (PR) is an algorithm used by Google search to rank web pages in their search engine results
  - PageRank is a way of measuring the importance of website pages



Graph-based ranking method

- TextRank



Topic-based clustering

- A document could cover different topics (e.g., sport, finance, …)
  - In graph-based methods, all the keyphrases could be selected from the same topic

- Here the idea is to grouping the candidate keyphrases in a document into topics
  1. A keyphrase should ideally be relevant to one or more main topic(s) discussed in a document
  2. The extracted keyphrases should be comprehensive in the sense that they should cover all the main topics in a document

Hasan, Kazi Saidul, and Vincent Ng. "Automatic keyphrase extraction: A survey of the state of the art." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1262-1273. 2014.

Topic-based clustering

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

Topic-based clustering

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

Topic #1	Topic #2
BERT	question answering
Bidirectional Encoder Representations	language inference
deep bidirectional representations	architecture modifications
output layer	

Supervised keyphrase selection

- The idea is to select the best keyphrases from the candidate list using labeled data (training data) for train a model
- Task reformulation
  - Binary classification
  - Ranking problem

Supervised keyphrase selection

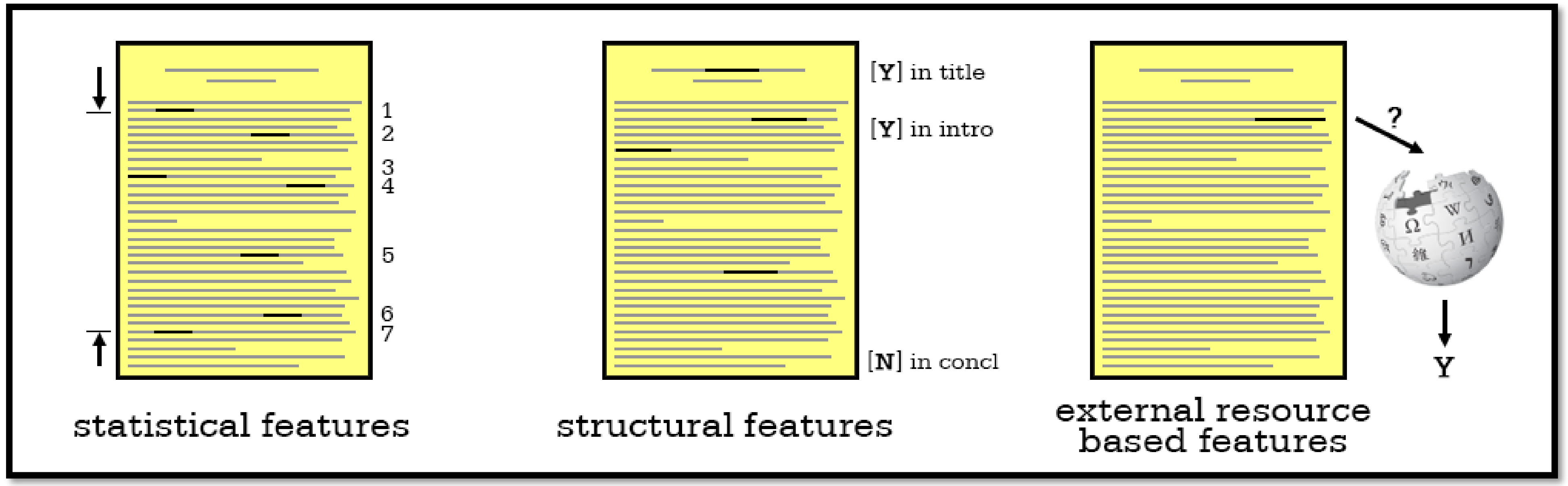
- Binary classification
  - Some fraction of candidates are classified as keyphrases and the rest as non-keyphrases
    - Train a classifier (Naive Bayes, SVM, …)
    - Label candidate keyphrases as True/False

- Ranking problem
  - We can also train a model to rank the candidate keyphrases instead of labeling them
  - And then choosing top N candidates from the ranked list

Hasan, Kazi Saidul, and Vincent Ng. "Automatic keyphrase extraction: A survey of the state of the art." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1262-1273. 2014.

Supervised keyphrase selection

- Common features to train a model
  - Phrase length (number of constituent words)
  - Phrase position (normalized position within a document)
  - Document's structural features (titles, abstracts, intros and conclusions, …)
    - A candidate is more likely to be a keyphrase if it appears in notable sections
  - Phrase commonness
    - Compares a candidate's frequency in a document with respect to its frequency in external corpora



https://bdewilde.github.io/

Hasan, Kazi Saidul, and Vincent Ng. "Automatic keyphrase extraction: A survey of the state of the art." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1262-1273. 2014.

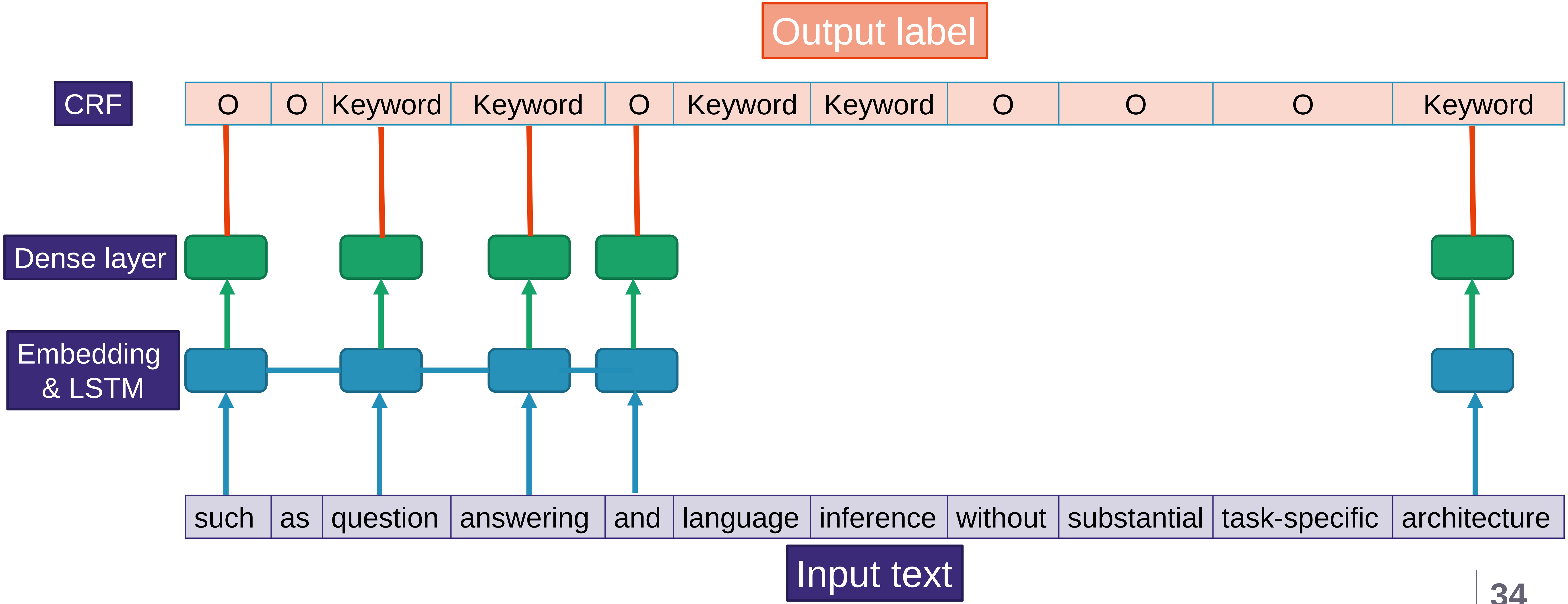
Keyphrase extraction

- What is keyphrase extraction
- Why is keyphrase extraction important
- Classical keyphrase extraction methods
- Neural keyphrase extraction
- Evaluation of automatic keyphrase extraction

Neural keyphrase extraction

- Feeding the input text to a neural network
- End to end approach
  - A machine learning model can directly convert an input data into an output prediction bypassing the intermediate steps that usually occur in a traditional pipeline
- Two common task formulations
  - Keyphrase extraction as sequence labeling
  - Keyphrase generation with sequence to sequence models

Neural keyphrase extraction



Deep keyphrase extraction

Language-specific Models in Multilingual Topic Tracking

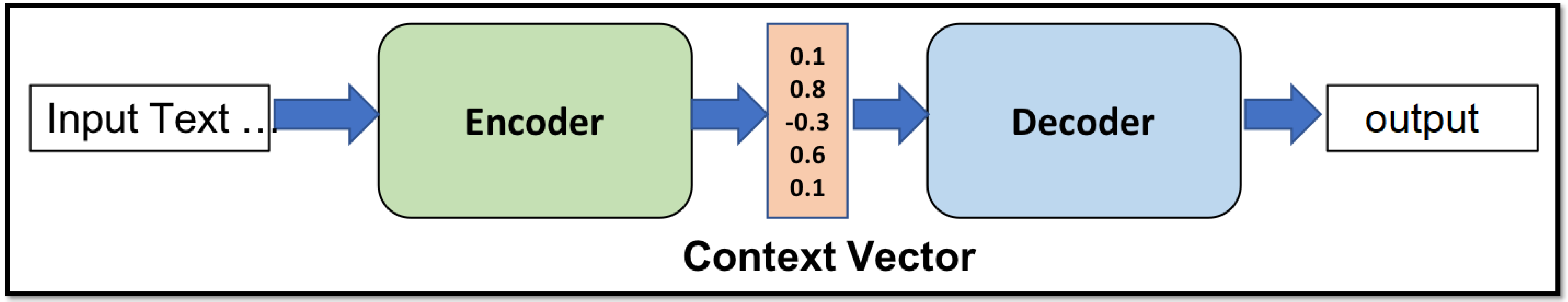
ABSTRACT
Topic tracking is complicated when the stories in the stream occur in multiple languages. Typically, researchers have trained only English topic models because the training stories have been provided in English. In tracking, non-English test stories are then machine translated into English to compare them with the topic models. We propose a *native language hypothesis* stating that comparisons would be more effective in the original language of the story. We first test and support the hypothesis for story link detection. For topic tracking the hypothesis implies that it should be preferable to build separate language-specific topic models for each language in the stream. We compare different methods of incrementally building such native language topic models.

Keywords: classification, crosslingual, Arabic, TDT, topic tracking, multilingual

Meng, Rui, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. "Deep keyphrase generation." arXiv preprint arXiv:1704.06879 (2017).

Neural keyphrase extraction

- From keyphrase extraction to generation
- sequence to sequence models



https://medium.com/nerd-for-tech

Keyphrase extraction

- What is keyphrase extraction
- Why is keyphrase extraction important
- Classical keyphrase extraction methods
- Neural keyphrase extraction
- Evaluation of automatic keyphrase extraction

Evaluation of automatic keyphrase extraction

- Keyphrase extraction
  - As a classification task
  - As a ranking task

Evaluation of automatic keyphrase extraction

- Keyphrase extraction as a classification task

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times (precision \times Recall)}{(precision + Recall)}$$

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: True Positive	FP: False Positive
	Negative	FN: False Negative	TN: True Negative

Evaluation of automatic keyphrase extraction

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: 6	FP: 4
	Negative	FN: 4	TN: -

Actual	Predicted
Deep learning	Confusion matrix
NLP	Train
Train	Model
Confusion matrix	Algorithm
Data	Data
Model	Validation
Machine learning	Test
Supervised	Classification
Clustering	Feed-forward
Classification	Supervised

Evaluation of automatic keyphrase extraction

- Keyphrase extraction as a classification task
  - Exact match is an overly strict condition, considering a predicted keyphrase incorrect even if it is a variant of the actual keyphrases
  - Confusion matrix → Confusion matrices
  - Neural network → neural net

- Common metrics from machine translation and text summarization reward a partial matches

- Same metrics can be used for the task of keyphrase extraction
  - BLEU, METEOR, and ROUGE

Evaluation of automatic keyphrase extraction

- Keyphrase extraction as a ranking task

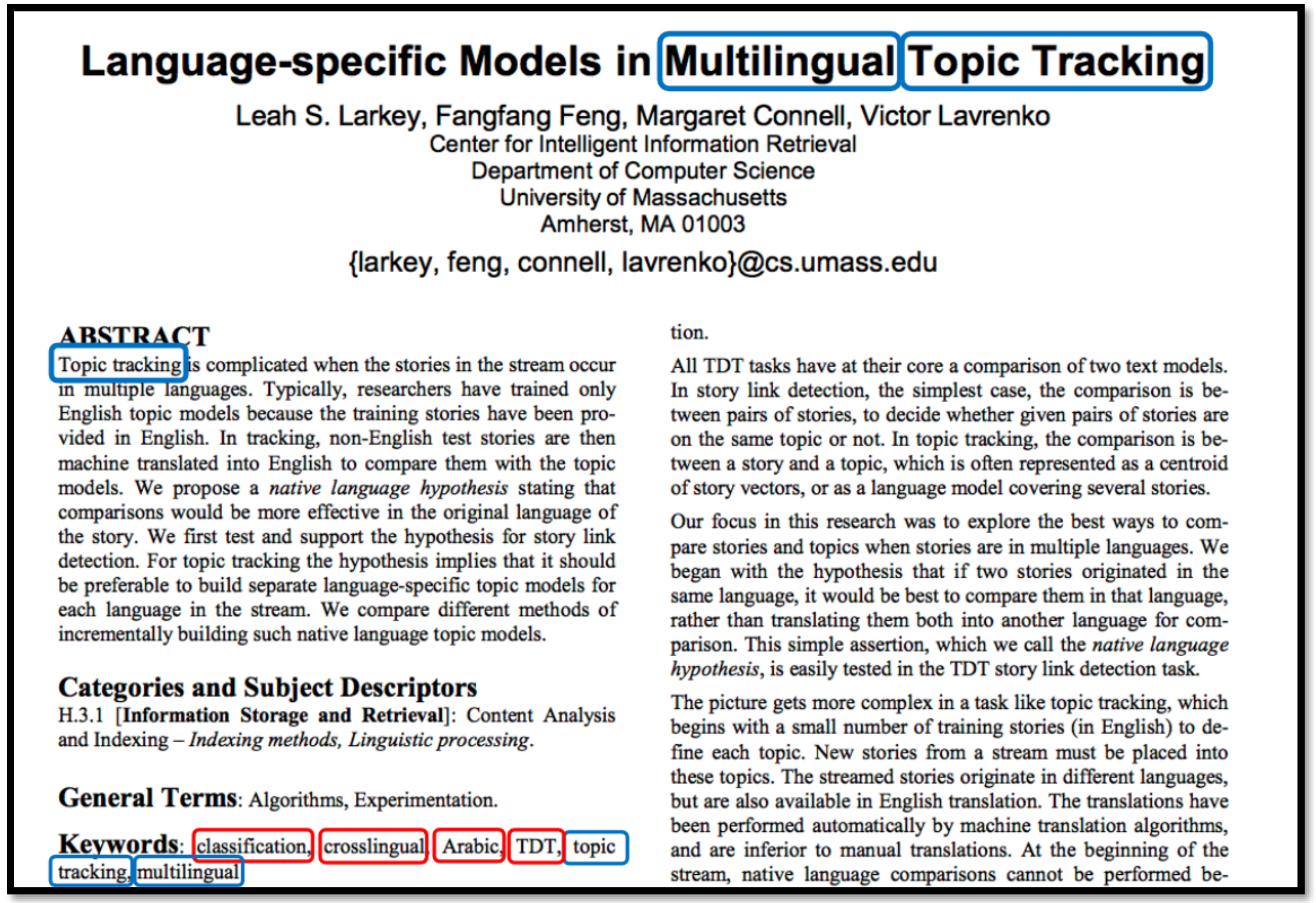
$$Precision@k = \frac{TP@k}{TP@k + FP@k}$$

Evaluation of automatic keyphrase extraction

Actual	Predicted	Precision@k
Deep learning	Confusion matrix	100%
NLP		
Train		
Confusion matrix		
Data		
Model		
Machine learning		
Supervised		
Clustering		
Classification		

Summary

- Keyphrase extraction is the automated process of extracting the most relevant words/phrases and expressions from text



Language-specific Models in Multilingual Topic Tracking

Leah S. Larkey, Fangfang Feng, Margaret Connell, Victor Lavrenko
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{larkey, feng, connell, lavrenko}@cs.umass.edu

ABSTRACT

Topic tracking is complicated when the stories in the stream occur in multiple languages. Typically, researchers have trained only English topic models because the training stories have been provided in English. In tracking, non-English test stories are then machine translated into English to compare them with the topic models. We propose a *native language hypothesis* stating that comparisons would be more effective in the original language of the story. We first test and support the hypothesis for story link detection. For topic tracking the hypothesis implies that it should be preferable to build separate language-specific topic models for each language in the stream. We compare different methods of incrementally building such native language topic models.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Indexing methods, Linguistic processing*.

General Terms: Algorithms, Experimentation.

Keywords: classification, crosslingual, Arabic, TDT, topic tracking, multilingual

tion.

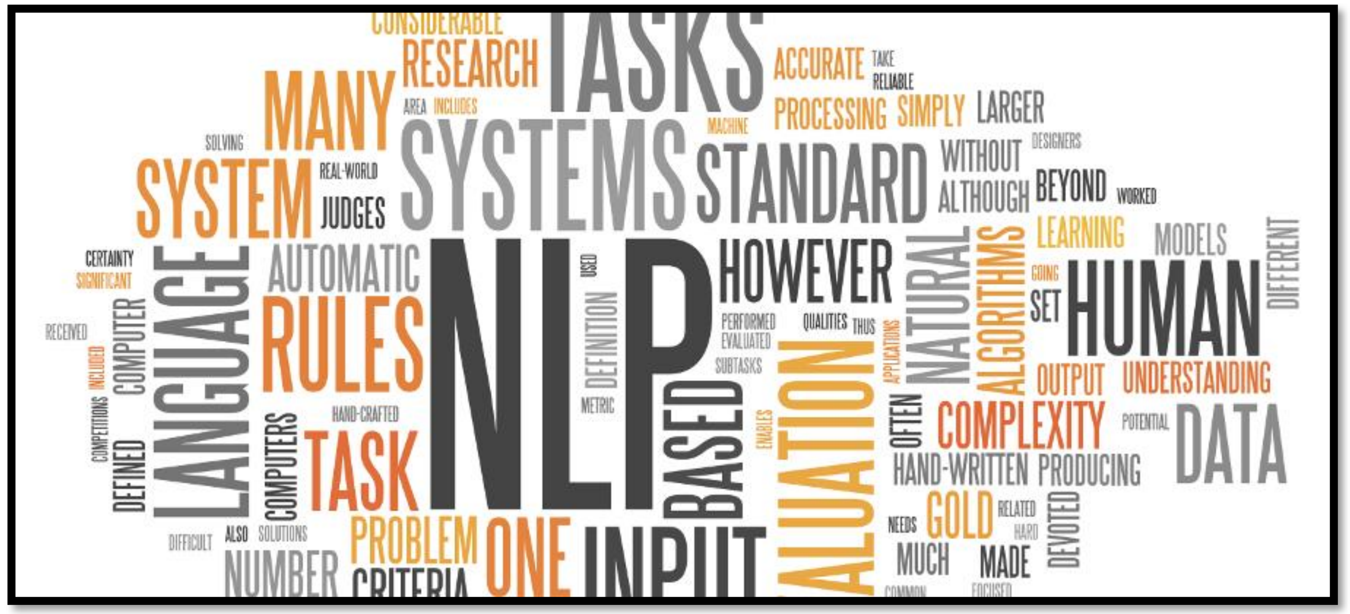
All TDT tasks have at their core a comparison of two text models. In story link detection, the simplest case, the comparison is between pairs of stories, to decide whether given pairs of stories are on the same topic or not. In topic tracking, the comparison is between a story and a topic, which is often represented as a centroid of story vectors, or as a language model covering several stories.

Our focus in this research was to explore the best ways to compare stories and topics when stories are in multiple languages. We began with the hypothesis that if two stories originated in the same language, it would be best to compare them in that language, rather than translating them both into another language for comparison. This simple assertion, which we call the *native language hypothesis*, is easily tested in the TDT story link detection task.

The picture gets more complex in a task like topic tracking, which begins with a small number of training stories (in English) to define each topic. New stories from a stream must be placed into these topics. The streamed stories originate in different languages, but are also available in English translation. The translations have been performed automatically by machine translation algorithms, and are inferior to manual translations. At the beginning of the stream, native language comparisons cannot be performed be-

Summary

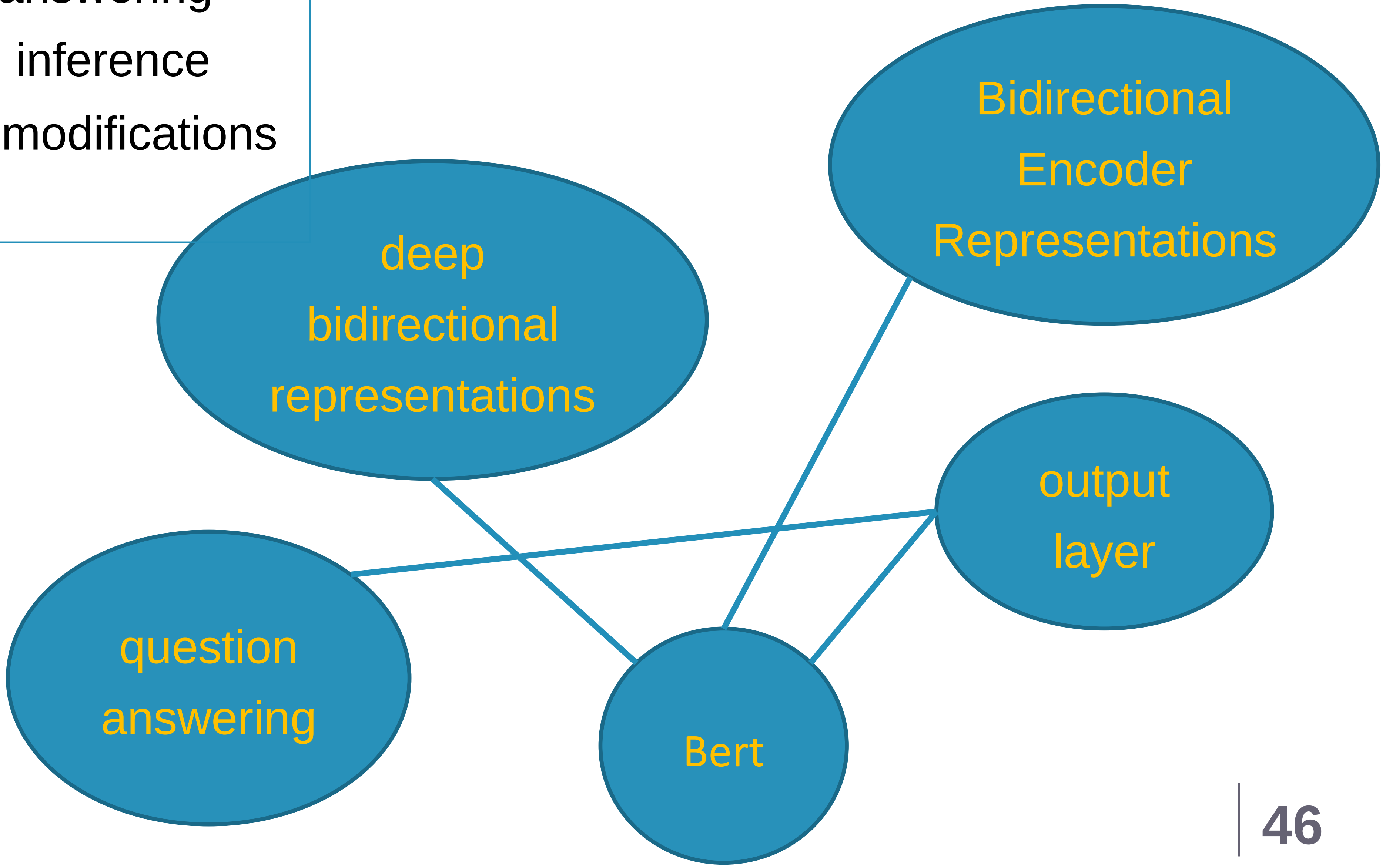
- Main applications
  - Automatic summarization
  - Text classification
  - Information visualization
  - …
- Classical keyphrase extraction
  - Candidate identification
  - Keyphrase selection
    - Unsupervised approaches
    - Supervised models



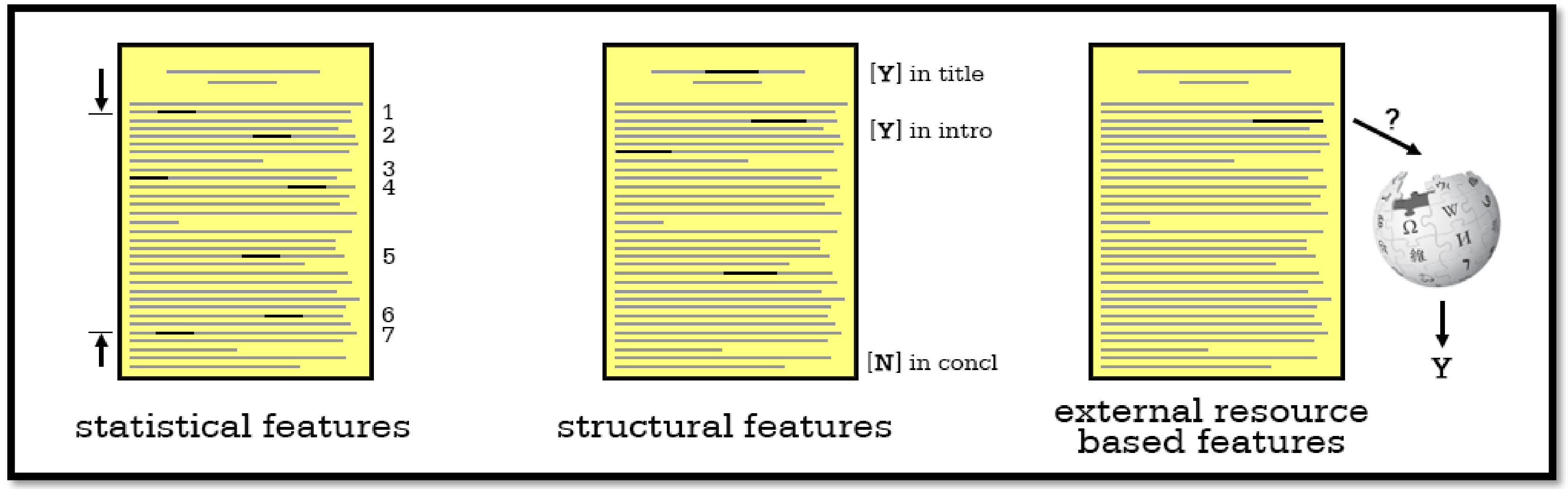
http://erikburger.nl

Summary

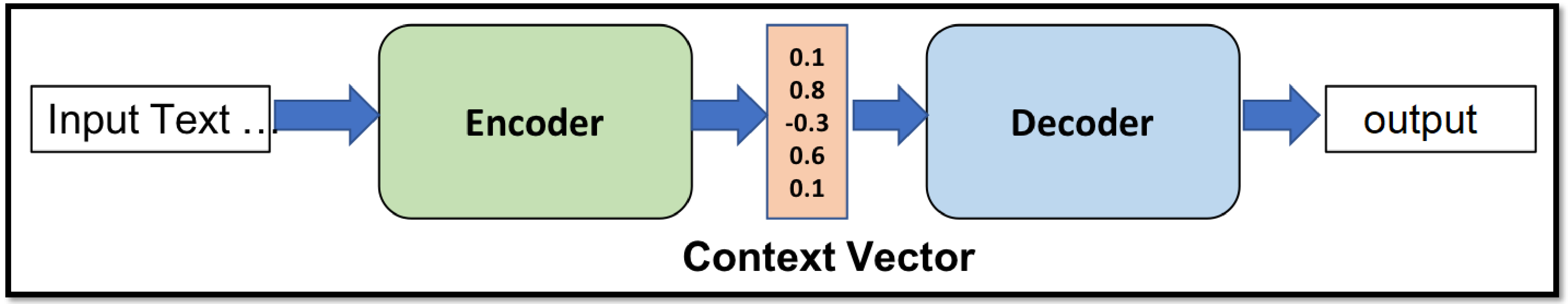
Topic #1	Topic #2
BERT	question answering
Bidirectional Encoder Representations	language inference
deep bidirectional representations	architecture modifications
output layer	



Summary



https://bdewilde.github.io/



Summary

<table>
<tr><td></td><td></td><td colspan="2">Actual class</td></tr>
<tr><td></td><td></td><td>Positive</td><td>Negative</td></tr>
<tr><td rowspan="2">Predicted class</td><td>Positive</td><td>TP: 6</td><td>FP: 4</td></tr>
<tr><td>Negative</td><td>FN: 4</td><td>TN: -</td></tr>
</table>

Actual	Predicted	Precision@k
Deep learning	Confusion matrix	100%
NLP	Train	100%
Train	Model	100%
Confusion matrix	Algorithm	75%
Data	Data	80%
Model	Validation	66%
Machine learning	Test	57%
Supervised	Classification	62%
Clustering	Feed-forward	55%
Classification	Supervised	60%