On the use of **Web resources** and **NLP techniques** to improve automatic **speech recognition** systems

Gwénolé Lecorvé, Guillaume Gravier, Pascale Sébillot

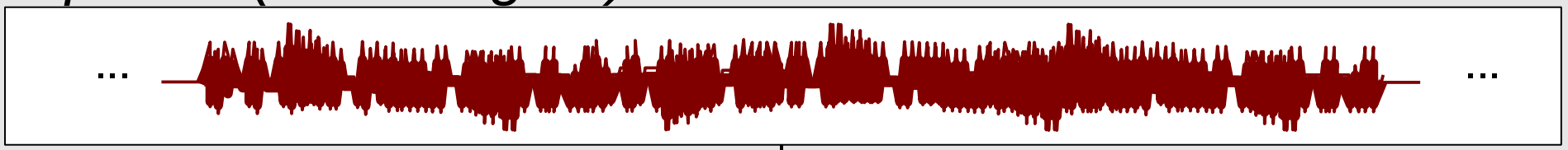
IRISA, France

IRISA

LREC 2008

Automatic Speech Recognition (ASR)

Speech (Audio signal)



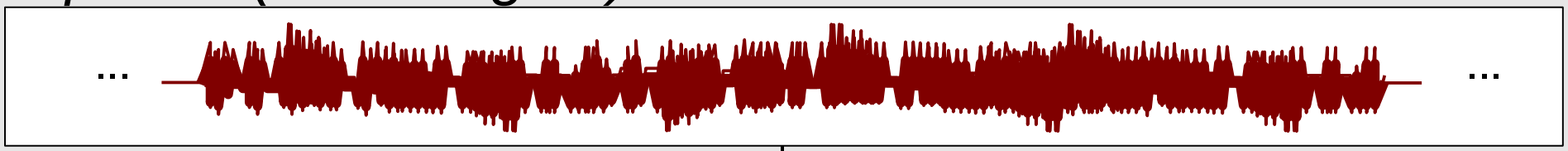
ASR system

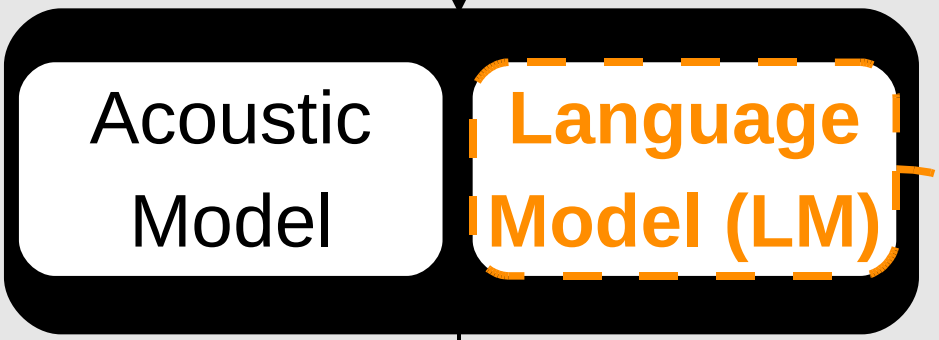
receives a single electoral boat in california

Transcript hypothesis (Text)

Automatic Speech Recognition (ASR)

Speech (Audio signal)





receives a single electoral **boat in california**

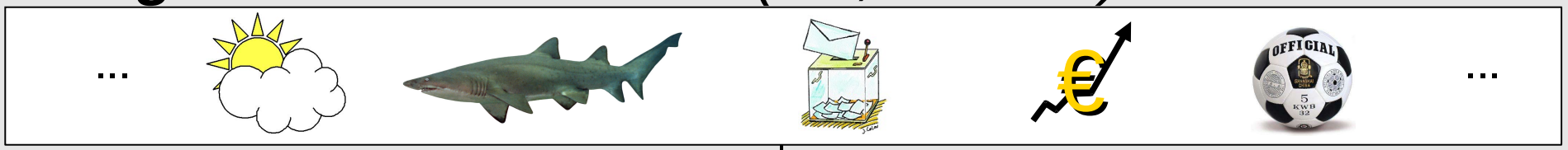
Transcript hypothesis (Text)

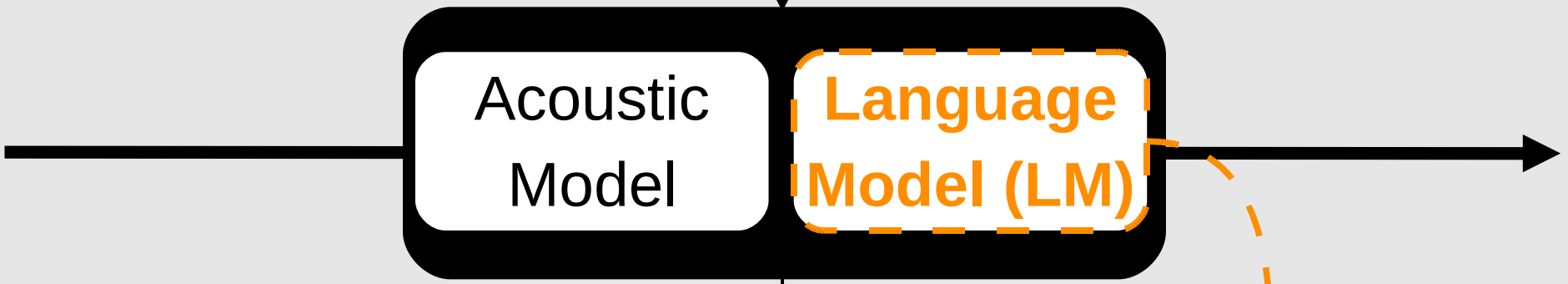
P(california | boat in)

N-gram probability

Automatic Speech Recognition (ASR)

Long multimedia stream *(i.e., 24H TV)*





`receives a single electoral` **`boat`** `in` `california`

Transcript hypothesis (Text)

P(california | boat in)

N-gram probability

Problem

Training process

- *N*-gram probabilities are trained
  - Once and for all
  - On a multi-topic collection of texts

Real conditions

- *N*-gram probabilities change according to the topic

Problem

Training process

- *N*-gram probabilities are trained
  - Once and for all
  - On a multi-topic collection of texts

Real conditions

- *N*-gram probabilities change according to the topic

→ **How to adapt a general-purpose LM**

- **to any topic**
- **with an unsupervised method?**

... using NLP techniques

Topic LM adaptation

- How to characterize the topic?
  - Pre-defined topics [Seymore and Rosenfeld 1997]
  - Discriminating words: TF-IDF [Suzuki 2006]

Topic LM adaptation

- How to characterize the topic?
  - ~~Pre-defined topics [Seymore and Rosenfeld 1997]~~
  - Discriminating words: TF-IDF [Suzuki 2006]

Topic LM adaptation

- How to characterize the topic?
  - ~~Pre-defined topics [Seymore and Rosenfeld 1997]~~
  - Discriminating words: TF-IDF [Suzuki 2006]

  → **Topic-specific texts collection**

Topic LM adaptation

- How to characterize the topic?
  - ~~Pre-defined topics [Seymore and Rosenfeld 1997]~~
  - Discriminating words: TF-IDF [Suzuki 2006]

  → **Topic-specific texts collection**

- How to select relevant texts?
  - Information Retrieval methods [Salton 1989]

Topic LM adaptation

- How to characterize the topic?
  - ~~Pre-defined topics [Seymore and Rosenfeld 1997]~~
  - Discriminating words: TF-IDF [Suzuki 2006]

  → **Topic-specific texts collection**

- How to select relevant texts?
  - Information Retrieval methods [Salton 1989]

- From which corpus?
  - From a static corpus [Klakow 2000]
  - From the Web [Berger and Miller 1998]

Topic LM adaptation

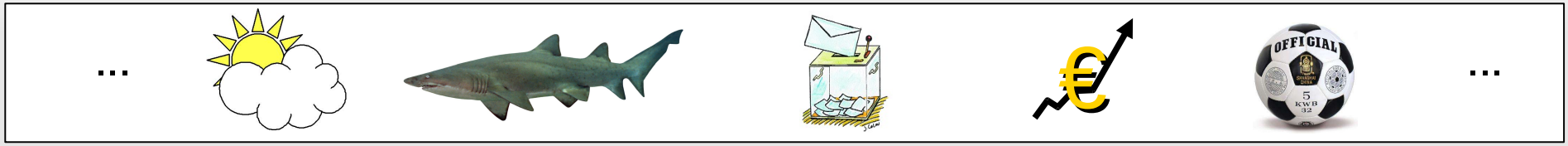
- How to characterize the topic?
  - ~~Pre-defined topics [Seymore and Rosenfeld 1997]~~
  - Discriminating words: TF-IDF [Suzuki 2006]

  → **Topic-specific texts collection**

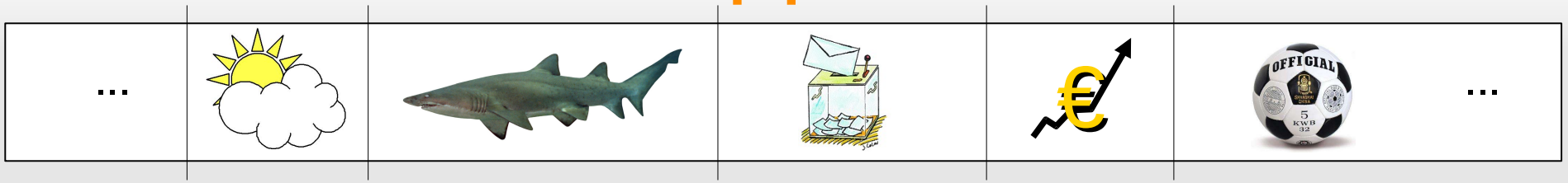
- How to select relevant texts?
  - Information Retrieval methods [Salton 1989]

- From which corpus?
  - ~~From a static corpus [Klakow 2000]~~
  - From the Web [Berger and Miller 1998]

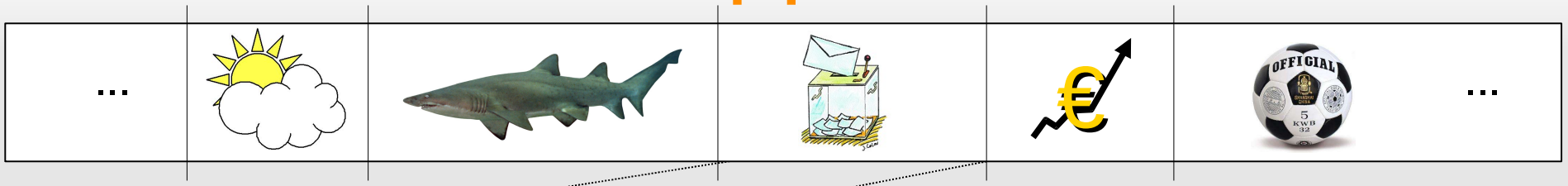
Our approach



Our approach



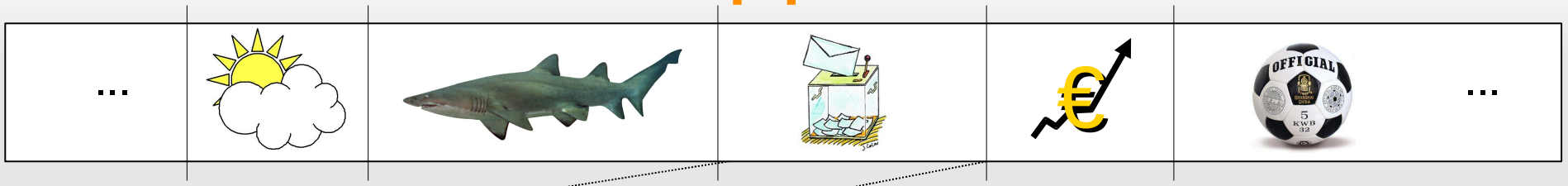
Our approach



…thus a candidate who fails to carry a particular state receives not a single electoral **boat** in that state for the popular votes received since residential elections are won by electoral ...

Baseline LM

Our approach

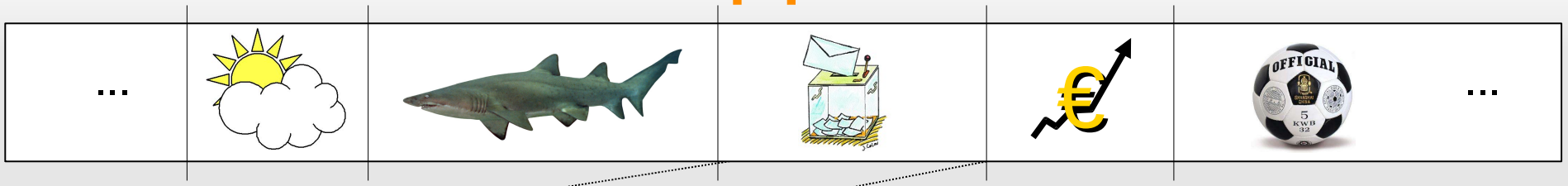


…thus a candidate who fails to carry a particular state receives not a single electoral **boat** in that state for the popular votes received since residential elections are won by electoral ...

Baseline LM

Adapted LM

Our approach



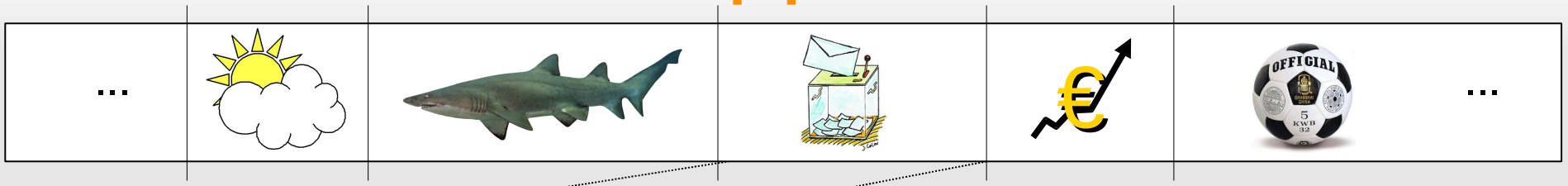
…thus a **candidate** who fails to carry a particular **state** receives not a single **electoral** **boat** in that **state** for the popular votes received since residential **elections** are won by **electoral** ...

1. Keyword extraction

Baseline LM

Adapted LM

Our approach



…thus a **candidate** who fails to carry a particular **state** receives not a single **electoral** **boat** in that **state** for the popular votes received since residential **elections** are won by **electoral** ...

1. Keyword extraction

`candidate state`

`electoral state`

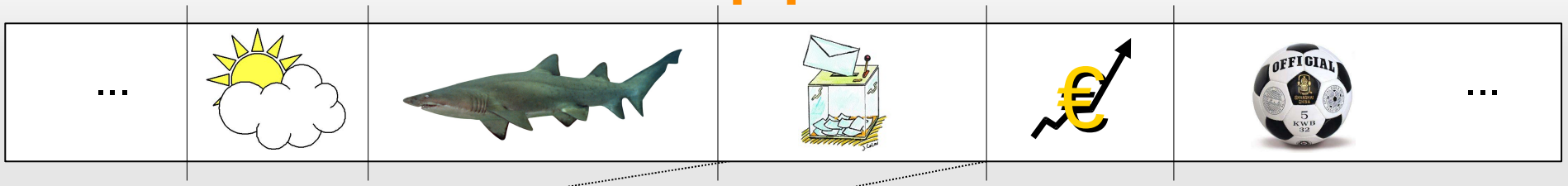
`candidate state election`

2. Querying

Baseline LM

Adapted LM

Our approach



…thus a **candidate** who fails to carry a particular **state** receives not a single **electoral** **boat** in that **state** for the popular votes received since residential **elections** are won by **electoral** ...

1. Keyword extraction

`candidate state`
`electoral state`
`candidate state election`

2. Querying

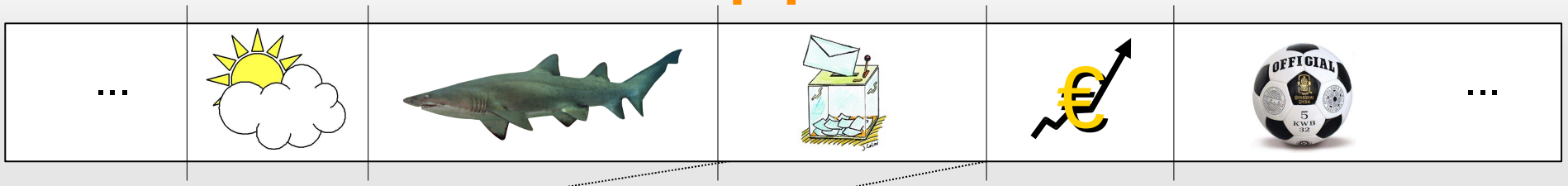
Web search engine

Baseline LM

Adapted LM

3. Retrieval of an adaptation corpus

Our approach



…thus a **candidate** who fails to carry a particular **state** receives not a single **electoral** **boat** in that **state** for the popular votes received since residential **elections** are won by **electoral** ...

1. Keyword extraction

```
candidate state
electoral state
candidate state election
```

2. Querying

Web search engine

3. Retrieval of an adaptation corpus

4.a Training of a topic-specific LM

4.b Mix of this LM and the general one

Adaptation LM + Baseline LM = Adapted LM

Keyword extraction

- Word scoring
  - High scores ↔ discriminative words
  - Score = TF-IDF + NLP modifications
- Enhancements using NLP
  - Lemmatization (reducing to a canonical form)
  - Penalty of proper names (POS tagging)
  - Acknowledgement of confidence measures

→ Sorted list of keywords

Querying

- Few first keywords

- Single complete query ?

  candidate boat state electoral election

  → NO

- Query-based sampling [Monroe et. al. 2002]

  candidate boat

  state election

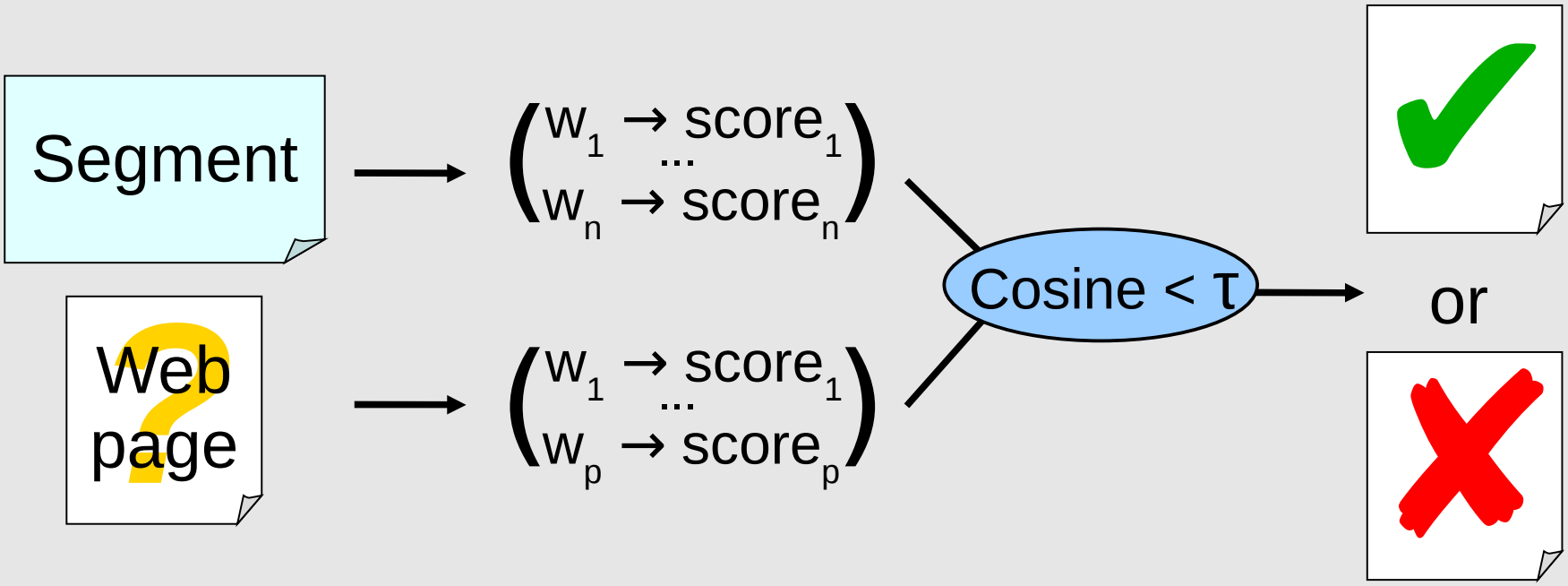
  ...

  candidate electoral election

  – Avoid mismatches due to misrecognized words

Creation of an adaptation corpus

- Filtering strategy
  - Number of Web pages retrieved = 200 pages
  - Thematic similarity = cosine distance



Results

- 2nd transcription pass using the adapted LMs
  - 172 segments ~ 6h radio broadcast news

	Word Error Rate	Lemma Error Rate
Baseline LM	21.7	19.6
Adapted LM	21.5	19.1
Difference	**-0.2**	**-0.5**

- Correction of “thematic terms”
- New grammatical errors

Conclusion & future work

- Use the Internet and natural language processing techniques
- Diagnose when topic adaptation is needed
  - Concept of topic
- Integrate more natural language processing techniques
  - Semantic links between keywords, complex terms
- Better use of thematic corpora
  - Bootstrap to precise the topic, adaptation of the vocabulary

Thank you

Questions ?

Adaptation corpus building 2/2

- No threshold
  - Average similarity / page = 0.08
  - Average similarity / corpus = 0.24

  } **A**

- Threshold = 0.08
  - Average similarity / selected page = 0.18
  - Average similarity / corpus = 0.35

  } **A**

} **B**

A : Query-based sampling effectiveness

B : Noise reducing thanks to the threshold

Results

Reference:	le	service	des	**trams**	est	**affecté**
	the	*tram*	*service*		*is*	*affected*
Baseline:	le	service	des	**trames**	est	**affectée**
	the	*woof*	*service*		*is*	*affected*
Adapted:	le	service	des	**tram_**	est	**affectée**
	the	*tram*	*service*		*is*	*affected*