SCIENCES SORBONNE UNIVERSITÉ

FROM NATURAL LANGUAGE TO KEYWORD QUERIES

Word Selection

6 mars 2019

Adrien Pouyet

LIP6

MLIA Machine Learning & Deep Learning for Information Access

Problem Description

Information Retrieval

SCIENCES SORBONNE UNIVERSITÉ

The idea

From a set of documents, retrieve the information the user wants.

- The user gives a query
- Find documents from your database to give the user informations

Example

- User : « Does climate change affect economics ? »
- Engine : « Yes ? No ? Look at those 1.000.000 documents I found »

Classical Model

Search engines are designed to work with isolated words (i.e. : keywords). The classical model is called BM25 and works as following :

$$score(D, Q) = \sum_{i=1}^{n} IDF(q_i) \frac{f(q_i, D)\dot{(}k_1 + 1)}{f(q_i, D) + k_1(1 - b + b\frac{|D|}{avgdl})}$$

Then, you just order documents according to their score and tada !

Problems

SCIENCES SORBONNE UNIVERSITÉ

User's problems

Formulating queries using keywords is not that easy. One might not know **the** keyword needed and will use multiple iterations to actually find what one is looking for.

Don't be fooled !

We want to give the user the information s·he's looking for, not the one s·he asked !

A complex task

In addition to retrieving documents, we have to understand between lines.

Classical Solutions

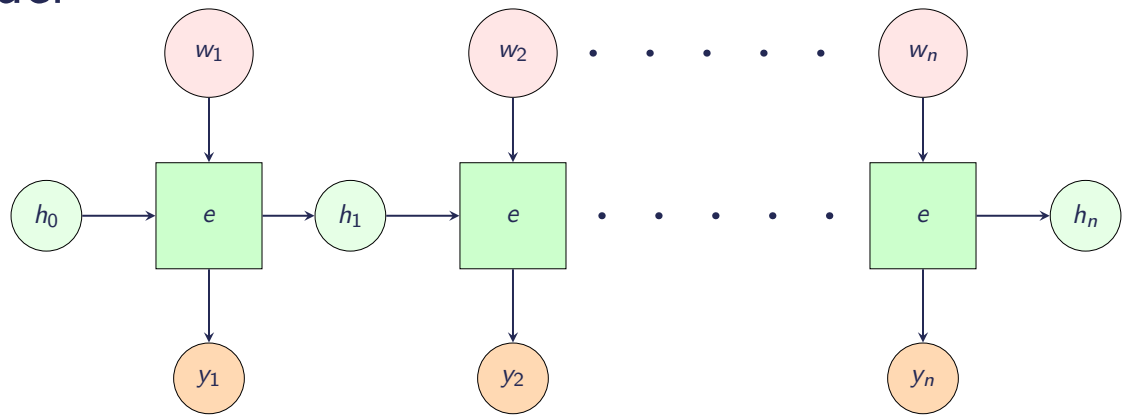
Usual answers to those problems are augmenting queries :

- Add every synonym found
- Extend the request with top documents titles
- Use the context (clicks, view times, ...) to have a better understanding of the user need (you can then weight the words)

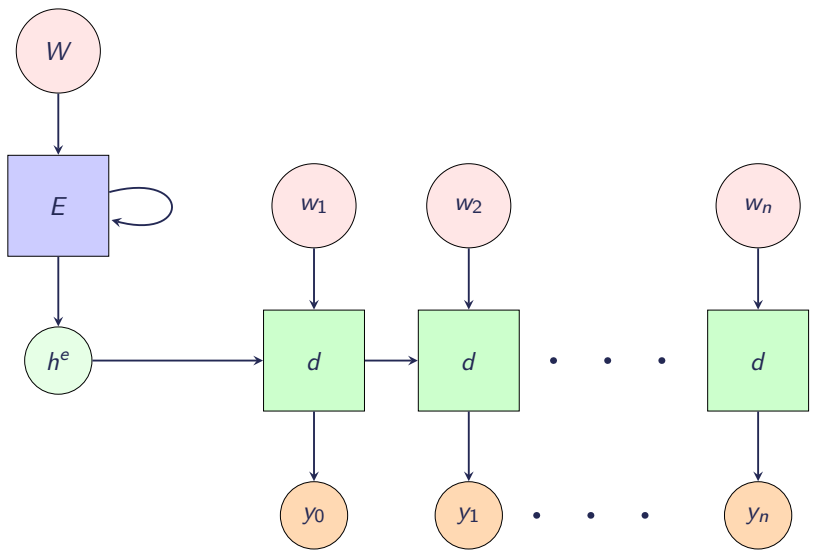
Our Approach

Select the Words

Basic encoder

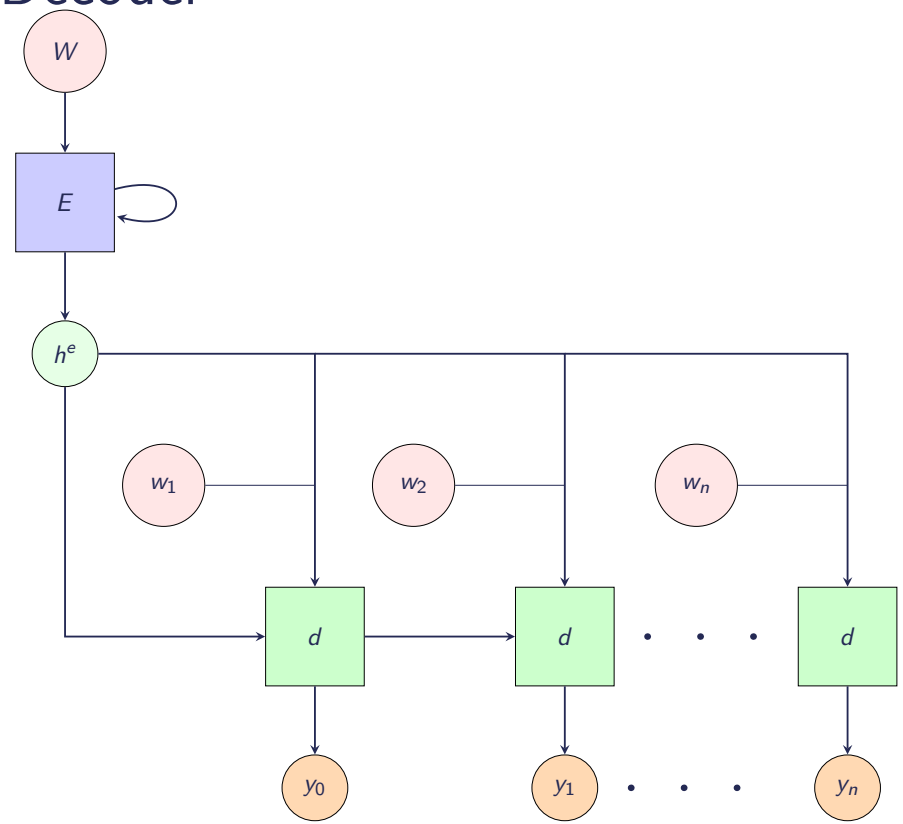


Encoder Decoder



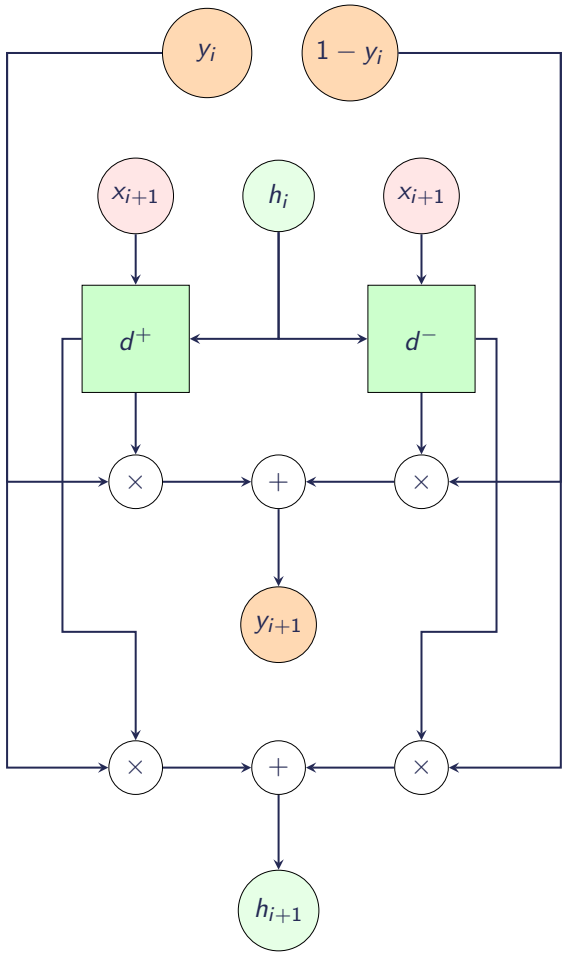
Select the Words

Mode Encoder Decoder

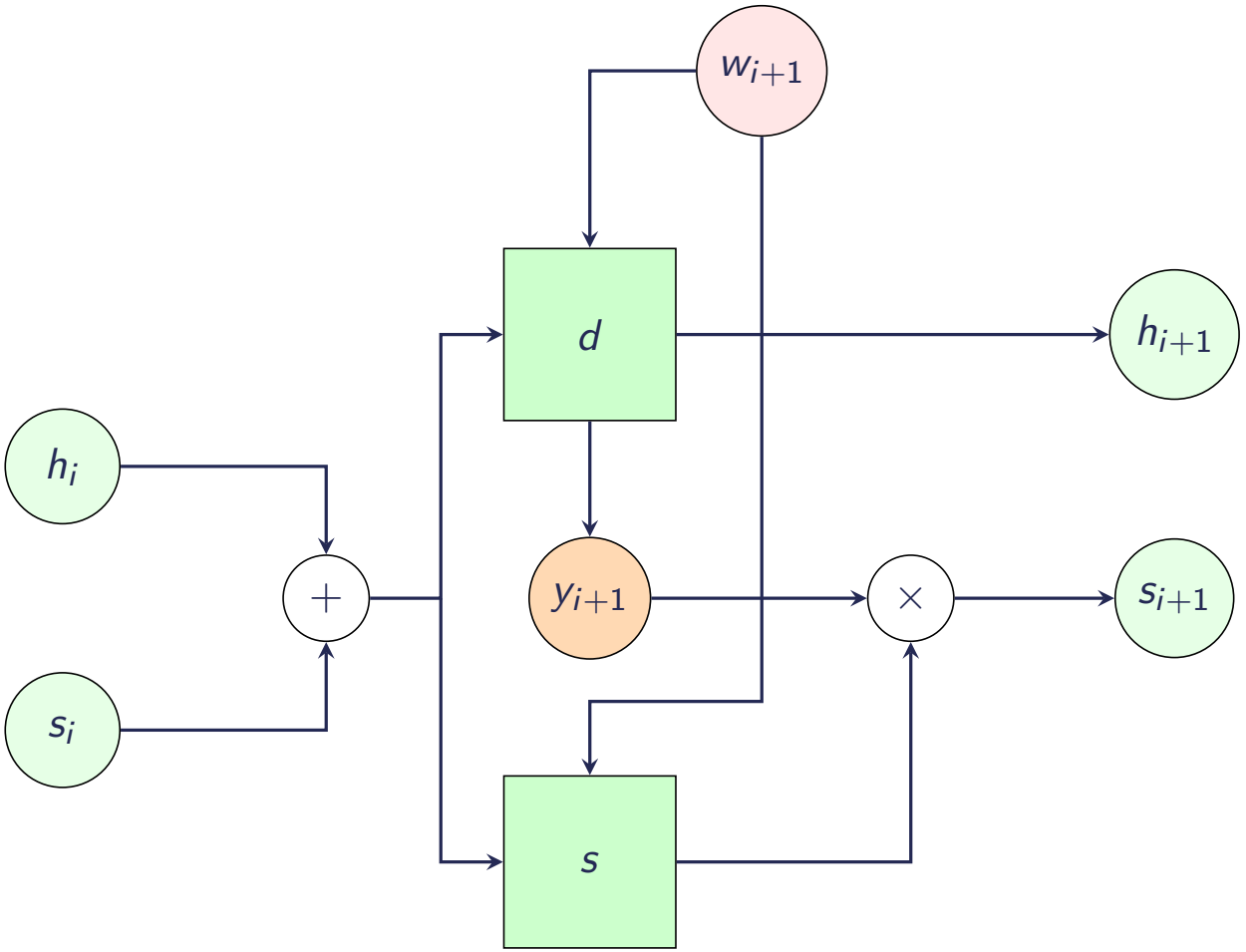


Memories !

Memory Switch



Hierarchical Memory



Results

Paper's Tab

Baseline	TREC Robust(2004)		TREC Web (2000-2001)	
	MAP	$\%Chg$	MAP	$\%Chg$
NL	0.08925	+15.25% ***	0.15913	+12.88% *
Q	0.09804	+4.92%	0.16543	+8.58%
Q bin	0.08847	+16.26% *	0.17402	+3.22%
Random	0.01808	+468.91% ***	0.04060	+342.44% ***
SMT	0.06845	+50.27% ***	0.08891	+102.04% ***
RL	0.08983	+14.51% ***	0.16474	+9.04%
SMT+RL	**0.10286**		**0.17963**	

Does it work really work ?

SCIENCES SORBONNE UNIVERSITÉ

Nope

Future Works

Future Works

Model Change

- Convolutions ?
- Transformers ?

Add Keyword Generation

- Try transfert learning with RL
- Deal with lack of supervision by restricting the language depending the query