27 Years of Text Mining: A Personal View on the Evolution of NLP, IR and ML over the last decades

E. Gaussier

Univ. Grenoble Alpes, CNRS, Grenoble INP – LIG
Skopai
Data Institute - Data@UGA

DS Meetup - 6. Dec. 2017

Table of Contents

Introduction to text mining

Text mining, also referred to as *text data mining*, is the process of deriving high-quality information from texts (Wikipedia, Dec. 2017).

What kind of information?

Ceci est du français	Fr
Now in English	En
Ora in italiano	It
Nio Grdink	Gb

Introduction to text mining

Text mining, also referred to as *text data mining*, is the process of deriving high-quality information from texts (Wikipedia, Dec. 2017).

What kind of information?

Ceci est du français	Fr
Now in English	En
Ora in italiano	It
Nio Grdink	Gb

Introduction to text mining (cont'd)

Typical applications (multilingual environments)

- Text categorization and clustering
- Topic, sentiment, viewpoint analysis
- Concept/entity extraction, relation extraction
- Document summarization

Underlying models and methods

- Natural language processing (NLP)
- Information retrieval (IR)
- Machine learning (ML)

Introduction to text mining (cont'd)

Typical applications (multilingual environments)

- Text categorization and clustering
- Topic, sentiment, viewpoint analysis
- Concept/entity extraction, relation extraction
- Document summarization

Underlying models and methods

- Natural language processing (NLP)
- Information retrieval (IR)
- Machine learning (ML)

Introduction to text mining (cont'd)

Data science approaches

- Text mining is data oriented: extracting information from *texts*
- IR/ML are also data oriented
  - Long empiricist tradition in IR
  - AI (knowledge-driven approaches) also present in ML
- NLP (computational linguistics)?
  - A long story in between computer science and linguistics
  - Numerical/statistical rise of the 80s-90s

Table of Contents

Statistical approaches to NLP

Machine translation as the prototypical example

"The attached memorandum on translation from one language to another, and on the possibility of contributing to this process by the use of modern computing devices of very high speed, capacity, and logical flexibility ..."

When was this written?

$\rightarrow$ In 1949, by Warren Weaver (1894, 1978)

"And it is one of the chief purposes of this memorandum to emphasize that statistical semantic studies should be undertaken, as a necessary preliminary step."

Translation, W. Weaver, 1949

Statistical approaches to NLP

Machine translation as the prototypical example

"The attached memorandum on translation from one language to another, and on the possibility of contributing to this process by the use of modern computing devices of very high speed, capacity, and logical flexibility ..."

When was this written?

→ In 1949, by Warren Weaver (1894, 1978)

"And it is one of the chief purposes of this memorandum to emphasize that statistical semantic studies should be undertaken, as a necessary preliminary step."

Translation, W. Weaver, 1949

Statistical approaches to NLP

Machine translation as the prototypical example

"The attached memorandum on translation from one language to another, and on the possibility of contributing to this process by the use of modern computing devices of very high speed, capacity, and logical flexibility ..."

When was this written?

$\rightarrow$ In 1949, by Warren Weaver (1894, 1978)

"And it is one of the chief purposes of this memorandum to emphasize that statistical semantic studies should be undertaken, as a necessary preliminary step."

Translation, W. Weaver, 1949

Statistical approaches to NLP

Machine translation as the prototypical example

"The attached memorandum on translation from one language to another, and on the possibility of contributing to this process by the use of modern computing devices of very high speed, capacity, and logical flexibility ..."

When was this written?

$\rightarrow$ In 1949, by Warren Weaver (1894, 1978)

"And it is one of the chief purposes of this memorandum to emphasize that statistical semantic studies should be undertaken, as a necessary preliminary step."

Translation, W. Weaver, 1949

Statistical approaches to NLP (cont'd)

Some key dates in the early development of machine translation (1)

1. 1954: Georgetown-IBM experiments (60 Russian sentences translated into English)
2. 1954-64: development of computational linguistics, boosted by the cold war
3. From J. A. Kouwenhaven, *The trouble with translation*, Harper's Magazine, 1962:
   - Translating from English to Russian, and back to English

   Out of sight, out of mind $\leftrightarrow$ Invisible idiot
4. 1964: ALPAC (Automatic Language Processing Advisory Committee) report - significant decrease in MT research in the US (not in other countries, incl. France)

J. Hutchins, *"The whisky was invisible", or Persistent myths of MT*, MT News International, 1995

Statistical approaches to NLP (cont'd)

Some key dates in the early development of machine translation (1)

1. 1954: Georgetown-IBM experiments (60 Russian sentences translated into English)
2. 1954-64: development of computational linguistics, boosted by the cold war
3. From J. A. Kouwenhaven, *The trouble with translation*, Harper's Magazine, 1962:
   - Translating from English to Russian, and back to English

     Out of sight, out of mind $\leftrightarrow$ Invisible idiot
4. 1964: ALPAC (Automatic Language Processing Advisory Committee) report - significant decrease in MT research in the US (not in other countries, incl. France)

J. Hutchins, *"The whisky was invisible", or Persistent myths of MT*, MT News International, 1995

Statistical approaches to NLP (cont'd)

Some key dates in the early development of machine translation (1)

1. 1954: Georgetown-IBM experiments (60 Russian sentences translated into English)
2. 1954-64: development of computational linguistics, boosted by the cold war
3. From J. A. Kouwenhaven, *The trouble with translation*, Harper's Magazine, 1962:
   - Translating from English to Russian, and back to English

     Out of sight, out of mind $\leftrightarrow$ Invisible idiot
4. 1964: ALPAC (Automatic Language Processing Advisory Committee) report - significant decrease in MT research in the US (not in other countries, incl. France)

J. Hutchins, *"The whisky was invisible", or Persistent myths of MT*, MT News International, 1995

Statistical approaches to NLP (cont'd)

Some key dates in the early development of machine translation (1)

1. 1954: Georgetown-IBM experiments (60 Russian sentences translated into English)
2. 1954-64: development of computational linguistics, boosted by the cold war
3. From J. A. Kouwenhaven, *The trouble with translation*, Harper's Magazine, 1962:
   - Translating from English to Russian, and back to English

     Out of sight, out of mind $\leftrightarrow$ Invisible idiot
4. 1964: ALPAC (Automatic Language Processing Advisory Committee) report - significant decrease in MT research in the US (not in other countries, incl. France)

J. Hutchins, *"The whisky was invisible", or Persistent myths of MT*, MT News International, 1995

Statistical approaches to NLP (cont'd)

Some key dates in the early development of machine translation (2)

5. 70s & 80s: statistical methods in speech recognition become mainstream
6. 1988: *A statistical approach to language translation*, P. Brown *et al.*, COLING 88 (paper by K. Church on stochastic parts program and noun phrase parser the same year)
7. 1993: *The mathematics of machine translation: parameter estimation*, P. Brown *et al.*, Comp. Linguistics Vol. 19(2)

Hansard parallel corpus - illustrative examples

The poor don't have money
Les pauvres sont démunis

No ifs, no buts, we want the truth
Pas de si, pas de mais, nous voulons la vérité

Statistical approaches to NLP (cont'd)

Some key dates in the early development of machine translation (2)

5. 70s & 80s: statistical methods in speech recognition become mainstream
6. 1988: *A statistical approach to language translation*, P. Brown *et al.*, COLING 88 (paper by K. Church on stochastic parts program and noun phrase parser the same year)
7. 1993: *The mathematics of machine translation: parameter estimation*, P. Brown *et al.*, Comp. Linguistics Vol. 19(2)

Hansard parallel corpus - illustrative examples

The poor don't have money
Les pauvres sont démunis

No ifs, no buts, we want the truth
Pas de si, pas de mais, nous voulons la vérité

Statistical approaches to NLP (cont'd)

Some key dates in the early development of machine translation (2)

5. 70s & 80s: statistical methods in speech recognition become mainstream
6. 1988: *A statistical approach to language translation*, P. Brown *et al.*, COLING 88 (paper by K. Church on stochastic parts program and noun phrase parser the same year)
7. 1993: *The mathematics of machine translation: parameter estimation*, P. Brown *et al.*, Comp. Linguistics Vol. 19(2)

Hansard parallel corpus - illustrative examples

The poor don't have money
Les pauvres sont démunis

No ifs, no buts, we want the truth
Pas de si, pas de mais, nous voulons la vérité

Statistical approaches to NLP (cont'd)

Many people participated to this effort at that time, on different aspects of NLP: P. Brown, B. Mercer, V. & S. Della Pietra, K. Church, W. Gale, M. Nagao, P. Isabelle, M. Simard, D. Wu, P. Fung, D. Melamed, B. Daille, E. Gaussier, ...

A lot of opposition: *The stone soup* (Y. Wilks, ?)

A few support from linguistics: distributional semantics (J. R. Firth, *you should know a word by the company it keeps*), development of corpus linguistics (J. Sinclair, Collins COBUILD)

The PCM (Pierce, Chomsky, Minsky) criticism

Statistical approaches to NLP (cont'd)

Some key dates in the early development of machine translation (3)

8. 1996: *A Maximum Entropy Approach to Natural Language Processing*, A. Berger *et al.*, Comp. Ling. Vol. 22(1)
9. Mid 90s/beg. 2000: rise of SVM (Support Vector Machines) based approaches (KerMIT project 2001-2004)
10. 2001: *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, Lafferty *et al.*, Int. Conf. on Machine Learning
11. 2004: *The Alignment Template Approach to Statistical Machine Translation*, F. J. Och & H. Ney, Comp. Ling., Vol. 30(4)
12. 2007: *A Pendulum Swung Too Far*, K. Church, Linguistic Issues in Language Technology, Vol. 2(4)
13. ca. 2010: the advent of deep learning (Y. Bengio, Y. Goldberg, L. Besacier, M. Dymetman, ... and many, many others)
14. 2017: *Neural Network Methods for Natural Language Processing*, Y. Goldberg, Morgan & Claypool (309 pages)

Statistical approaches to NLP (cont'd)

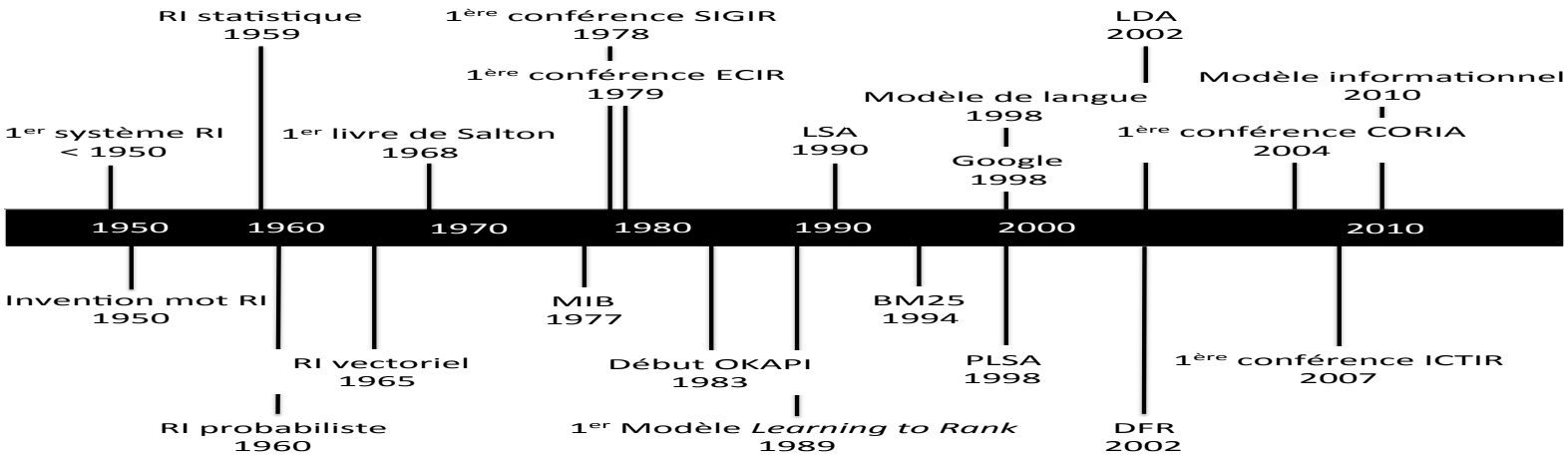
Text generation (E2E NLG challenge)

name[The Eagle], eatType[coffee shop], priceRange[moderate], kidsFriendly[yes]
$\rightarrow$ *The coffee shop, The Eagle, gives families a mid-priced dining experience.*

Table of Contents

What happened in IR during the last decades? (1)

- IR has been data-driven from the beginning, statistical/probabilistic models playing an important role from the 60s on (K. Spärck Jones, S. Robertson)
- In the early 90s, probabilistic models become predominant (over vector space and boolean models); already a ranking problem
- A true search engine on the Web in 1998 (Google)



From Amini & Gaussier, *Recherche d'Information: applications, modèles et algorithmes*, Eyrolles, 2013

What happened in IR during the last decades? (2)

The Web is not a standard collection!

Idea 1 Index (increased) web pages with terms in anchor texts

Idea 2 If two pages are similar content-wise, rank higher the one that is more *important* in the graph of the Web (*PageRank*)

What happened in IR during the last decades? (2)

The Web is not a standard collection!

Idea 1 Index (increased) web pages with terms in anchor texts

Idea 2 If two pages are similar content-wise, rank higher the one that is more *important* in the graph of the Web (*PageRank*)

What happened in IR during the last decades? (2)

The Web is not a standard collection!

Idea 1 Index (increased) web pages with terms in anchor texts

Idea 2 If two pages are similar content-wise, rank higher the one that is more *important* in the graph of the Web (*PageRank*)

What happened in IR during the last decades? (3)

Importance of queries (years 2000)

1. With Web search engines, possibility to collect *many* queries with click information, thus possibility to get relative relevance judgements
2. $\Rightarrow$ Learning to rank
   - Look for a function f that preserves partial order bet. docs (for a given query): $x_{(i)} \prec x_{(j)} \Longleftrightarrow f(x_{(i)}) < f(x_{(j)})$, with $x_{(i)}$ being a (query,doc) pair: $x_i = (d_i, q)$
   - Idea Transform a ranking information into a classification information by forming the difference between pairs:

$$x^{(i,j)} = (x_i - x_j, z = \begin{cases} +1 \text{ if } x_i \prec x_j \\ -1 \text{ if } x_j \prec x_i \end{cases})$$

   - Apply standard classification techniques on the above training set!

T. Qin *et al.*, *LETOR: A Benchmark Collection for Research on Learning to Rank for Information Retrieval*, Information Retrieval Journal, 2010

What happened in IR during the last decades? (4)

Importance of deep learning (years 2010)

Latent topic models vs word embeddings
(*topical representations vs semantic representations*)

Latent topic model

$$P(d = "w_1^d, ..., w_l^d") = \sum_{t=1}^{T} P(t)P(("w_1^d, ..., w_l^d"|t)$$
$$\approx \sum_{t=1}^{T} P(t)\pi_{w \in V} P(w|t)^{\#(w,d)}$$

- One obtains latent representations for words through: $P(t|w), 1 \leq t \leq T$ (T-dimensional word embedding)
- à la Firth (*company of a word*), but in a loose way (cooccurrences at the document level)

What happened in IR during the last decades? (5)

Importance of deep learning (years 2010)

Latent topic models vs word embeddings
(*topical representations vs semantic representations*)

Skip-gram model

$$\sum_{-c \leq j \leq c} logP(w_{i+j}|w_i)$$

- One looks for T-dimensional vectors that maximize the above quantity
- Still à la Firth, but with very limited context (more likely to correspond to semantic dimensions)
- *The representation is learned so as to optimize he objective function*

Kenter *et al.*, *Neural Networks for Information Retrieval*, tutorial SIGIR 2017

Table of Contents

What happened in ML during the last decades? (1)

- Many developments we have seen are rooted in ML (currently, no clear separations bet. communities)
- 1990s: shift from knowledge-driven to data-driven (SVMs, recurrent neural networks (RNNs))
- 2000s: kernel methods
- 2010s: deep learning (word embeddings, convolution neural networks, recurrent neural networks)

Short (simplified) history of ML (Wikpedia, Dec. 2017); several theoretical developments not mentioned here

What happened in ML during the last decades? (2)

From feature engineering to feature learning

- Feature engineering: features are "manually" designed (usually by experts); time consuming, difficult task

- Feature (representation) learning: representations are automatically learned from the data, for a specific objective function (dictionary learning, matrix factorization, metric learning, auto-encoders, neural networks (deep learning))

What happened in ML during the last decades? (3)

Why is deep learning so successful?

An important success: image classification with deep convolutional networks – Krizhevsky *et al.*, *ImageNet Classification with Deep Convolutional Neural Networks* NIPS 2012 (G. Hinton)

- Learning representation is key to many applications
- *Reality is non-linear* – Non-linear functions, universal approximation; complex (number of parameters) models that can also "learn by heart" in some cases
- Can make use of lots of data through efficient learning (GPUs)
- But:
  1. May require a lot of engineering (tweaking)
  2. Complex models require lots of training data
  3. Non-convex models: dependence on intialization and local optima
  4. Lack of clear theoretical understanding (S. Mallat)

Table of Contents

Putting it all together

- Several models (in particular RNNs - J. Schmidhuber) bring answers to the PCM criticism
- Does it mean that we are done?
  - Need to better understand the behaviours and limitations of current deep models
  - Still far from capturing subtle phenomena(?)

No ifs, no buts, we want the truth
Cessons de tourner autour du pot ... de chambre

Putting it all together

- Several models (in particular RNNs - J. Schmidhuber) bring answers to the PCM criticism
- Does it mean that we are done?
  - Need to better understand the behaviours and limitations of current deep models
  - Still far from capturing subtle phenomena(?)

No ifs, no buts, we want the truth

Cessons de tourner autour du pot ... de chambre

Putting it all together

- Several models (in particular RNNs - J. Schmidhuber) bring answers to the PCM criticism
- Does it mean that we are done?
  - Need to better understand the behaviours and limitations of current deep models
  - Still far from capturing subtle phenomena(?)

No ifs, no buts, we want the truth
Cessons de tourner autour du pot ... de chambre