Topic Models and Word Embeddings

Simón Roca <sroca@ing.uc3m.es>
Machine Learning Group - UC3M
17/05/2017

Contenidos

1. Topic Models
2. Word Embeddings
3. Similitudes y diferencias
4. Comparativa de rendimiento
5. Modelos híbridos
6. Conclusiones y líneas futuras
7. Referencias

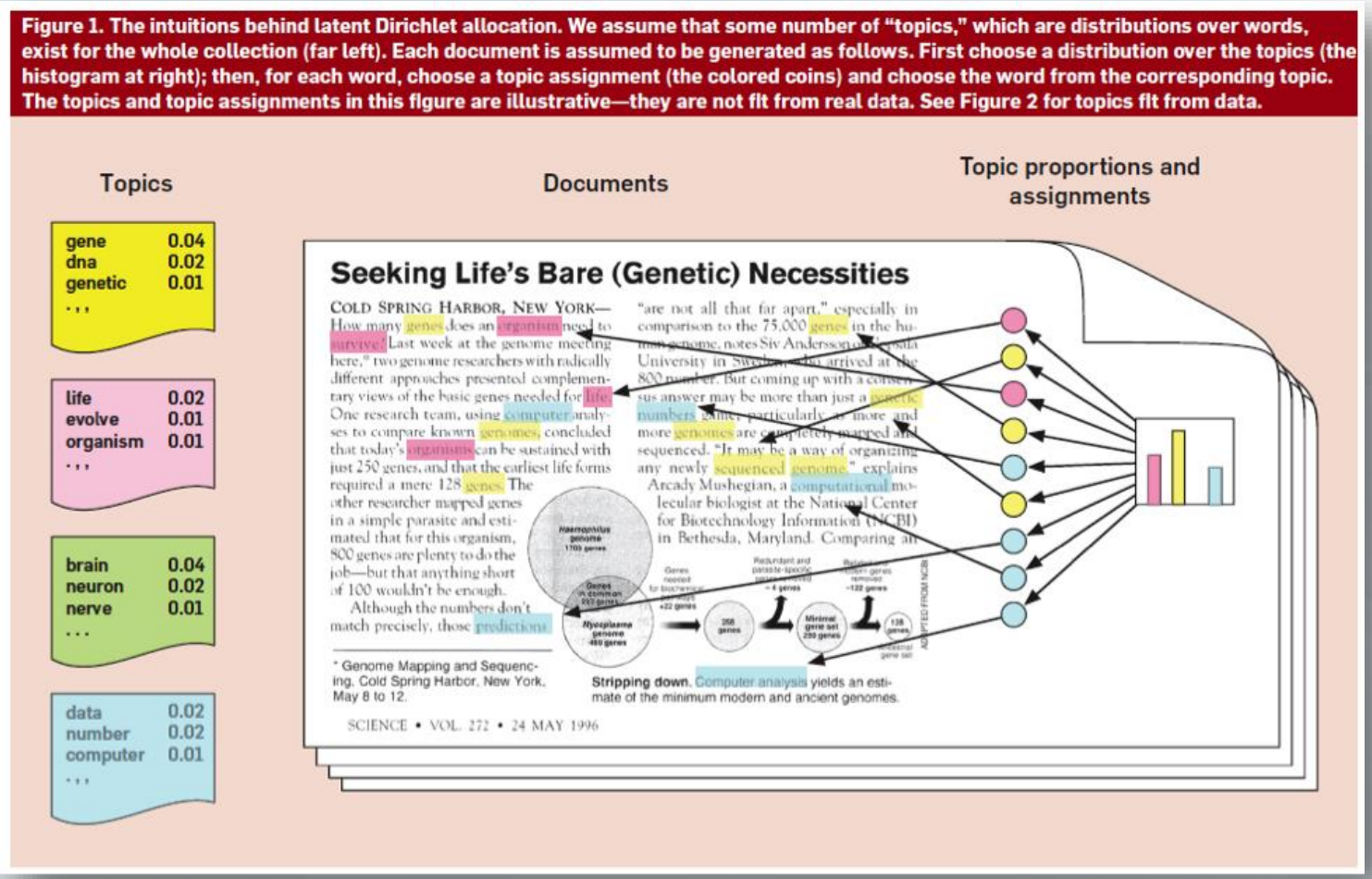
1. Topic Models

- Para una colección de documentos:
  - Descubrir tópicos automáticamente (no supervisado).
  - Tópico (LDA): distribución de probabilidades de aparición de distintas palabras del vocabulario.
  - Caracterizar un documento como mezcla de tópicos gracias a modelo de variables latentes.

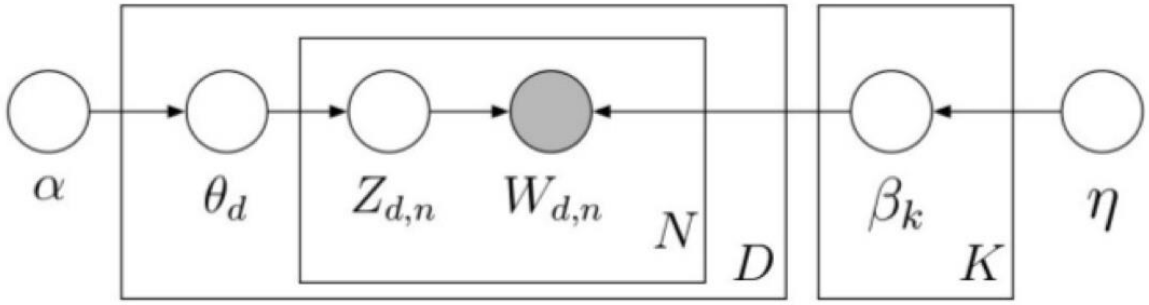
1. Topic Models

- Breve cronología:
  - 1986: Deerwester, S., et al, Improving Information Retrieval with Latent Semantic Indexing → **(LSI)**
  - 1999: Thomas Hofmann, Probabilistic Latent Semantic Indexing → **(pLSI)**
  - 2003: Blei, David M.; Ng, Andrew Y.; Jordan, Michael I. «Latent Dirichlet allocation» → **(LDA)**

1. Topic Models



1. Topic Models



K – total number of topics
β_k – topic, a distribution over the vocabulary
D – total number of documents
Θ_d – per-document topic proportions
N – total number of words in a document (it fact, it should be N_d)
$Z_{d,n}$ – per-word topic assignment
$W_{d,n}$ – observed word
α, η – Dirichlet parameters

- Several **inference algorithms** are available (e.g. sampling based)
- A few **extensions** to LDA were created:
  - Bigram Topic Model

$$p(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K} \mid w_{1:D,1:N}, \alpha, \eta) = \frac{p(\vec{\theta}_{1:D}, \vec{z}_{1:D}, \vec{\beta}_{1:K} \mid \vec{w}_{1:D}, \alpha, \eta)}{\int_{\vec{\beta}_{1:K}} \int_{\vec{\theta}_{1:D}} \sum_{\vec{z}} p(\vec{\theta}_{1:D}, \vec{z}_{1:D}, \vec{\beta}_{1:K} \mid \vec{w}_{1:D}, \alpha, \eta)}.$$

1. Topic Models

- Extensiones:
  - Supervisado: sLDA, discLDA, LLDA…
  - Atributos a los tópicos: hLDA, CTM, HLLDA…
  - Atributos a los docs: relational TM, dynTM, DMR…

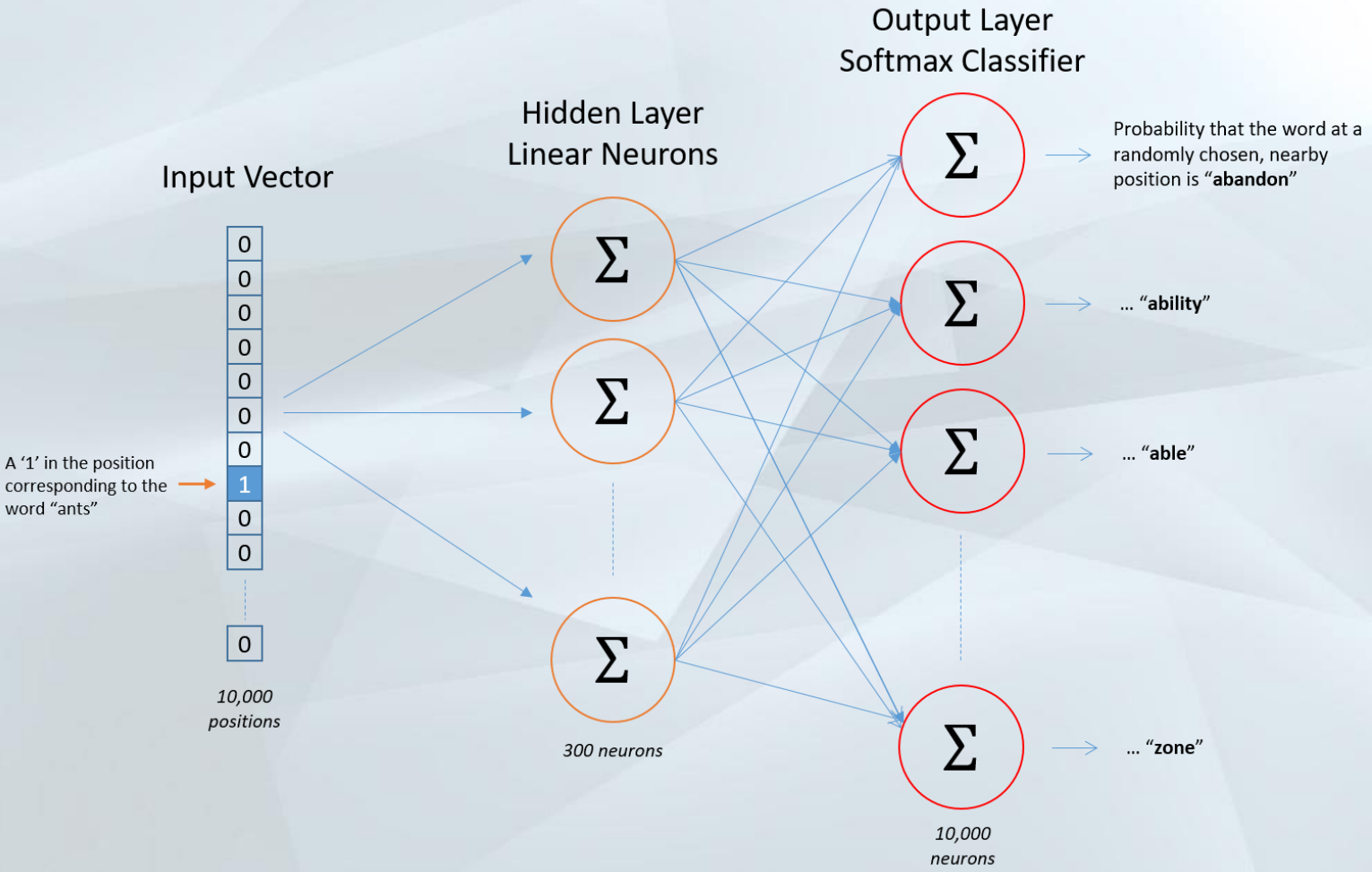
2. Word embeddings

- Para un vocabulario:
  - Representación real, densa y de baja dimensionalidad de palabras, explorando contexto local de las mismas.
  - (word2vec) Se consigue con una red neuronal que modela la probabilidad de que una palabra tenga en su contexto cercano cualquiera de las otras del vocabulario.

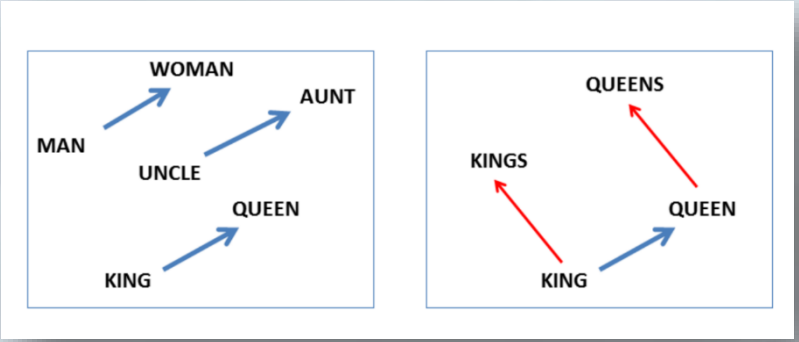
2. Word embeddings

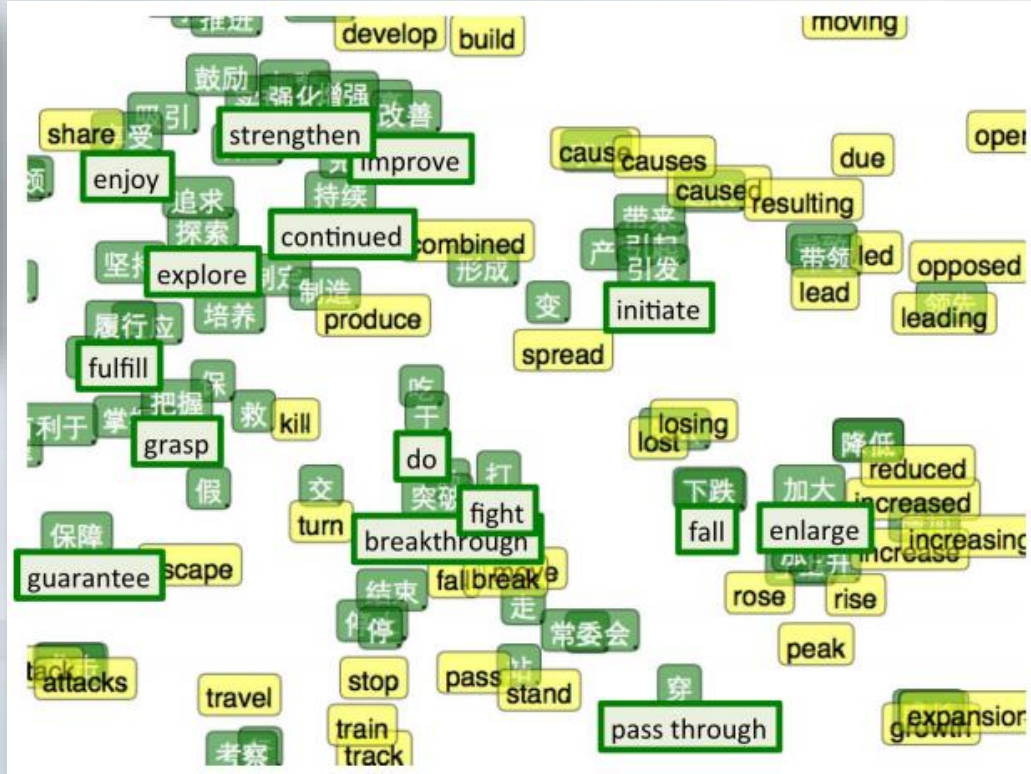
- Breve cronología:
  - “A word is characterized by the company it keeps” 1957 Firth, J.R. "A synopsis of linguistic theory 1930-1955“
  - 2003 Bengio et al. “Neural Probabilistic Language Models”
  - 2013 Mikolov et al. “Distributed representations of words and phrases and their compositionality”

2. Word embeddings



2. Word embeddings





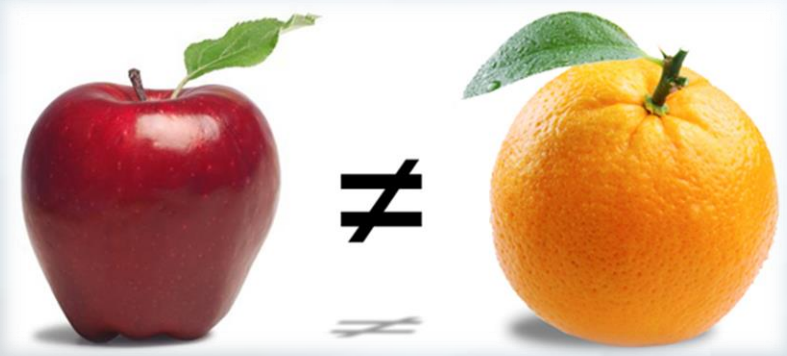
2. Word embeddings

- Extensiones y/o implementaciones:
  - Word2vec
  - FastText (word2vec + n-grams)
  - GloVe
  - Lexvec
  - …

3. Similitudes y diferencias

- Similitudes:
  - Modelos probabilísticos.
  - Capturan relaciones semánticas.
  - No supervisados.
- Diferencias:
  - Finalidad (contexto global vs. contexto local).
  - Dimensionalidad.

4. Comparativa de rendimiento



Si bien existe comparativas (no concluyentes) entre métodos de “conteo de palabras” (Bag of Words) y métodos “predictivos” (Word embeddings), son diferencias basadas en la representación de las palabras.

Como tal, aunque LDA use BoW, su finalidad es inferir tópicos y representar documentos, y las palabras como realizaciones de los mismos.

A día de hoy no se ha aceptado por consenso una superioridad entre una u otra representación.

4. Comparativa de rendimiento

- Aplicaciones
  - Topic models:
    - Clustering en general
    - Matching
    - Traducción
    - Visualización
    - Representación y similitud de documentos
    - Relación y evolución de tópicos (*)
  - Word embeddings:
    - Representación y similitud de palabras
    - Ambigüedad (*)
    - Traducción
    - Analogías
    - Sentiment Analysis
    - Machine comprehension

5. Modelos híbridos

- GaussianLDA
- LDA2VEC
- Mejoras en:
  - Coherencia de tópicos.
  - Document clustering y document classification.
  - Word relatedness. (car – road)

6. Conclusiones y líneas futuras

- Topic models y word embeddings capturan relaciones semánticas, pero difieren en implementación y utilización.
- WE es popular y en algunos experimentos sobresale frente a BoW; en otros (Levy et al.) se asocia la mejora a hiperparámetros.
- No hay consenso en la migración a WE.
- Se buscan modelos híbridos que sean capaces de capturar contexto local (palabras) y global (documentos).
- Dichos modelos deberían permitir visualizar los datos y a su vez ser útiles representaciones para problemas posteriores.

7. Referencias

- 2003: Blei, David M.; Ng, Andrew Y.; Jordan, Michael I. “Latent Dirichlet allocation”
- 2013: Mikolov et al., “Distributed Representations of Words and Phrases and their Compositionality”
- 2016: Lin Liu et al., “An overview of topic modeling and its current applications in bioinformatics”
- 2016: Jordy VL, “A Survey of Word Embedding Literature: Context Representations and the Challenge of Ambiguity”
- … and a big etc.