Application of Variable Length *N*-gram Vectors to Monolingual and Bilingual Information Retrieval

Daniel Gayo Avello

(University of Oviedo)

Introduction

- *blindLight* is a modified vector model with applications to several NLP tasks.
- Goals for this experience:
  - Test the application of *blindLight* to IR (monolingual).
  - Present a simple technique to pseudo-translate query vectors to perform bilingual IR.
- First group participation in CLEF tasks:
  - Monolingual IR (Russian)
  - Bilingual IR (Spanish-English)

Vector Model vs. *blindLight* Model

What's *n*-gram significance?

- **Can we know how important an *n* gram is within just one document without regards to any external collection?**
- Similar problem: Extracting multiword items from text (e.g. **`European Union`**, **`Mickey Mouse`**, **`Cross Language Evaluation Forum`**).
- Solution by Ferreira da Silva and Pereira Lopes:
  - Several statistical measures generalized to be applied to arbitrary length **word** *n*-grams.
  - New measure: **Symmetrical Conditional Probability** (SCP) which outperforms the others.
- So, our proposal to **answer first question**:

If **SCP** shows the most significant multiword items within just one document it **can be applied to rank character *n* gams for a document according to their significances.**

27%

What's *n*-gram significance? (cont.)

- Equations for SCP:

$$Avp = \frac{1}{n-1}\sum_{i=1}^{i=n-1} p(w_1 \ldots w_i)\cdot p(w_{i+1} \ldots w_n)$$

$$SCP\_f((w_1 \ldots w_n)) = \frac{p(w_1 \ldots w_n)^2}{Avp}$$

- **($w_1 \ldots w_n$)** is an *n*-gram. Let's suppose we use quad-grams and let's take **`(igni)`** from the text **`What's n-gram significance.`**
  - **($w_1 \ldots w_1$)** / **($w_2 \ldots w_4$)** = **`(i)`** / **`(gni)`** ①
  - **($w_1 \ldots w_2$)** / **($w_3 \ldots w_4$)** = **`(ig)`** / **`(ni)`**
  - **($w_1 \ldots w_3$)** / **($w_4 \ldots w_4$)** = **`(ign)`** / **`(i)`** ②
  - For instance, in ① **p(($w_1 \ldots w_1$))** = **p(`(i)`)** would be computed from the relative frequency of appearance within the document of *n*-grams starting with **`i`** (e.g. **`(igni)`**, **`(ific)`**, or **`(ican)`**).
  - In ② **p(($w_4 \ldots w_4$))** = **p(`(i)`)** would be computed from the relative frequency of appearance within the document of *n*-grams ending with **`i`** (e.g. **`(m_si)`**, **`(igni)`**, or **`(nifi)`**).

What's *n*-gram significance? (cont.)

- **Current implementation of *blindLight* uses quad-grams because…**
  - They provide better results than tri-grams.
  - Their significances are computed faster than $n \geq 5$ *n*-grams.
- **¿How would it work mixing different length *n*-grams within the same document vector?** Interesting question to solve in the future…
- **Two example *blindLight* document vectors:**
  - **Q document:** **`Cuando despertó, el dinosaurio todavía estaba allí.`**
  - **T document:** **`Quando acordou, o dinosauro ainda estava lá.`**
  - **Q vector** (45 elements):
    `{(Cuan, 2.49), (l_di, 2.39), (stab, 2.39), ..., (saur, 2.31), (desp, 2.31), ..., (ando, 2.01), (avía, 1.95), (_all, 1.92)}`
  - **T vector** (39 elements):
    `{(va_l, 2.55), (rdou, 2.32), (stav, 2.32), ..., (saur, 2.24), (noss, 2.18), ..., (auro, 1.91), (ando, 1.88), (do_a, 1.77)}`
- **¿How can such vectors be numerically compared?**

Comparing *blindLight* doc vectors

- **Comparing different length vectors is similar to pairwise alignment** (e.g. Levenshtein distance).
- **Levenshtein distance:** number of insertions / deletions / substitutions to change one string into another one.
- Some **examples**:
  - **Bioinformatics**: **`AAGTGCCTATCA`** vs. **`GATACCAAATCATGA`** (distance: 8)
    - **`AAG--TGCCTA-TCA---`**
    - **`---GATACCAAATCATGA`**
  - **Natural language:** Q document vs. T document (distance: 23)
    - **`Cuando_desper-tó,_el_dino-saurio_todavía_estaba_allí.`**
    - **`Quando_--acordou,_-o_dinossaur-o_--ainda_estava_--lá.`**
- **Relevant differences between text strings and *blindLight* doc vectors make sequence alignment algorithms not suitable:**
  - Doc vectors have term weights, strings don't.
  - Order of characters ("terms") within strings is important but unimportant for *bL* doc vectors.
- **Anyway… Sequence alignment has been inspiring…**

Comparing *blindLight* doc vectors (cont.)

- Some equations:

Comparing *blindLight* doc vectors (cont.)

$$Pi = S_{Q\Omega T}/S_Q = 20.48/97.52 = 0.21$$

$$Rho = S_{Q\Omega T}/S_T = 20.48/81.92 = 0.25$$

Information Retrieval using *blindLight*

- **Π (Pi) and P (Rho) can be linearly combined into different association measures to perform IR.**
- Just two tested up to now: Π and $\frac{\Pi + norm(\Pi P)}{2}$ (which performs slightly better).
- IR with *blindLight* is pretty easy:
  1. For **each document** within the **datas** ... nd stored.
  2. When a **query** is submitted to the syst ...
     a) A **4-gram (Q)** is computed for the qu ...
     b) For **each doc vector (T)**:
        i. Q and T are Ω-intersected obtainin ...
        ii. Π and P are combined into a uniqu ... .
     c) A **reverse ordered list of documents** is built and returned to **answer** the query.

Rho, and thus Pi ·Rho, values are negligible when compared to Pi. *norm* function scales Pi ·Rho values into the range of Pi values.

- **Features and issues:**
  – No indexing phase. Documents can be added at any moment. ☺
  – Comparing each query with every document not really feasible with big data sets. ☹

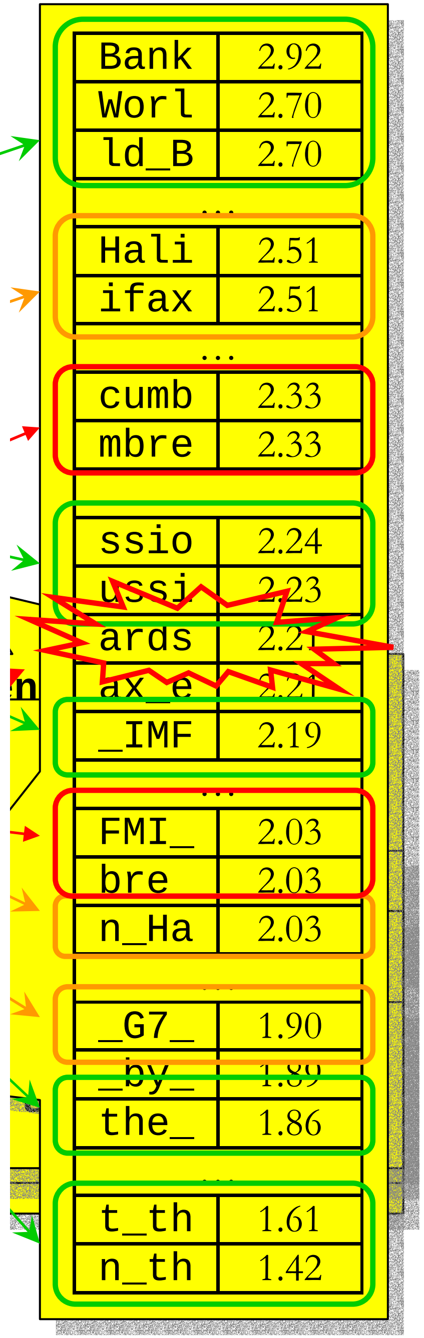
Bilingual IR with *blindLight*

We have compared *n*-gram vectors for pseudo-translations with vectors for actual translations (Source: Spanish, Target: English).

38.59% of the *n*-grams within pseudo-translated vectors are also within actual translations vectors.

28.31% of the *n*-grams within actual translations vectors are present at pseudo-translated ones.

Promising technique but thorough work is required.



Results and Conclusions

- **Monolingual IR within Russian documents:**
  - 72 documents found from 123 relevant ones.
  - Average precision: 0.14
- **Bilingual IR using Spanish to query English docs:**
  - 145 documents found from 375 relevant ones.
  - Average precision: 0.06.
- **Results at CLEF are far, far away from good but anyway encouraging… Why?!**
  - Quick and dirty prototype. *"Just to see if it works…"*
  - First participation in CLEF.
  - Not all topics achieve bad results. **THE PROBLEM** are mainly **broad topics** (e.g. **`Sportswomen and doping`**, **`Seal-hunting`**, **`New Political Parties`**).
  - Similarity measures are not yet tuned. Could genetic programming be helpful? Maybe…
- **To sum up, *blindLight*…**
  - **…is an extremely simple technique.**
  - **…can be used to perform information retrieval** (among other NLP tasks).
  - **…allows us to provide bilingual information retrieval trivially by performing "pseudo-translation" of queries.**

Application of Variable Length *N*-gram Vectors to Monolingual and Bilingual Information Retrieval

Daniel Gayo Avello
(University of Oviedo)

Danke schön

Ευχαριστώ

Thank you

Kiitos

Merci

Grazie

ありがとう

Dank u

Obrigado

Спасибо

Tack

谢谢

Gracias