Bilingual Document Alignment with Latent Semantic Indexing

Ulrich Germann, University of Edinburgh

The Bilingual Document Alignment Task

Given a collection of web page downloads and their corresponding ULRs, find pairs of pages that are translations of one another. A seed set of page pairs is given for development purposes.

Basic Approach

Compute the cosine between **document embeddings** in a joint semantic space to measure cross-lingual document similarity. We use *Latent Semantic Indexing* (LSI) to map documents into the joint space. LSI relies on reduced-rank *Singular Value Decomposition* (SVD) to perform this task.

$$\text{terms}\left\{\begin{array}{c|cccc} & \overbrace{d_1 \quad d_2 \quad \dots \quad d_m}^{\text{documents}} & & & \\ \hline t_1 & 1^a & 0 & \dots & 1 \\ t_2 & 0 & 1 & \dots & 0 \\ t_3 & 1 & 1 & \dots & 1 \\ t_4 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_n & 0 & 1 & \dots & 1 \end{array}\right. \approx \begin{array}{c|ccc} t'_1 & \cdot & \cdot & \cdot \\ t'_2 & \cdot & \cdot & \cdot \\ t'_3 & \cdot & \cdot & \cdot \\ t'_4 & \cdot & \cdot & \cdot \\ \vdots & \cdot & \cdot & \cdot \\ t'_n & \cdot & \cdot & \cdot \end{array} \quad \underbrace{\begin{array}{ccc} \cdot & 0 & 0 \\ 0 & \cdot & 0 \\ 0 & 0 & \cdot \end{array}}_{\text{Eigen values}} \quad \begin{array}{cccc} d'_1 & d'_2 & \dots & d'_m \\ \hline \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{array}$$

$$\mathbf{M} \qquad \mathbf{T} \qquad \mathbf{S} \qquad \mathbf{D}^{\mathsf{T}}$$

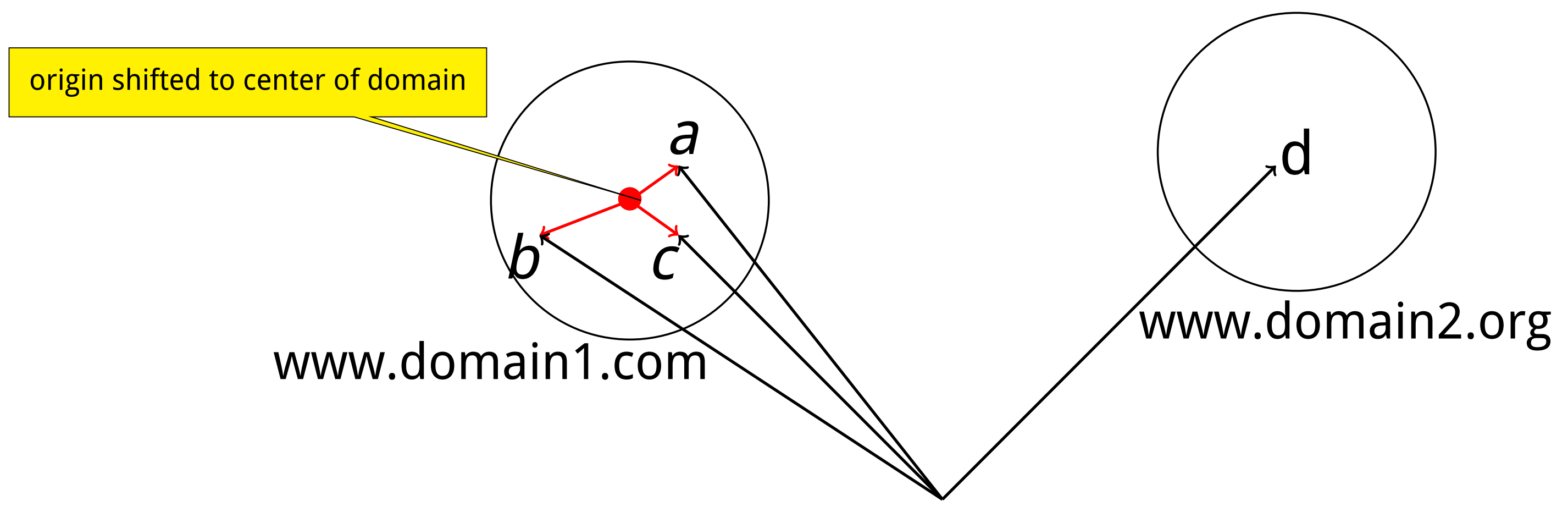
[a] in practice, we use tf-idf weighted term counts; see below.

Technical Procedure

1. Fill the term-document matrix with $tf_{t,d} \cdot idf_{t,D_d}$ values for each term occurring in each document pair d from the training data, with
$$tf_{t,d} = 1 + log(f_{t,d})$$
$$idf_{t,D_d} = \log \frac{|D_d|}{|\{d' \in D : t \in d'\}|}$$
where $f_{t,d}$ is the raw occurrence count of term t in d and D_d the web domain that document d belongs to.
2. Factorize term-document matrix with SVD.
3. For each document in the data set, compute a document vector $\vec{d}$ as in Step 1.
4. Fold it in: $\vec{d}' = S^{-1}\, T^{\mathsf{T}} \vec{d}$.
5. Compute document similarity between two documents d_x an d_y as
$$sim(d_x, d_y) = \cos(S\vec{d}'_x, S\vec{d}'_y).$$

Features

- URL string similarity (details in the paper)
- Cosine in the global space (cos)
- Cosine after shifting origin to center of documents from a specific web site (lcos) by subtracting the mean document vector over all documents from that site from each individual vector



Results (rank=1000)

recall on the training data

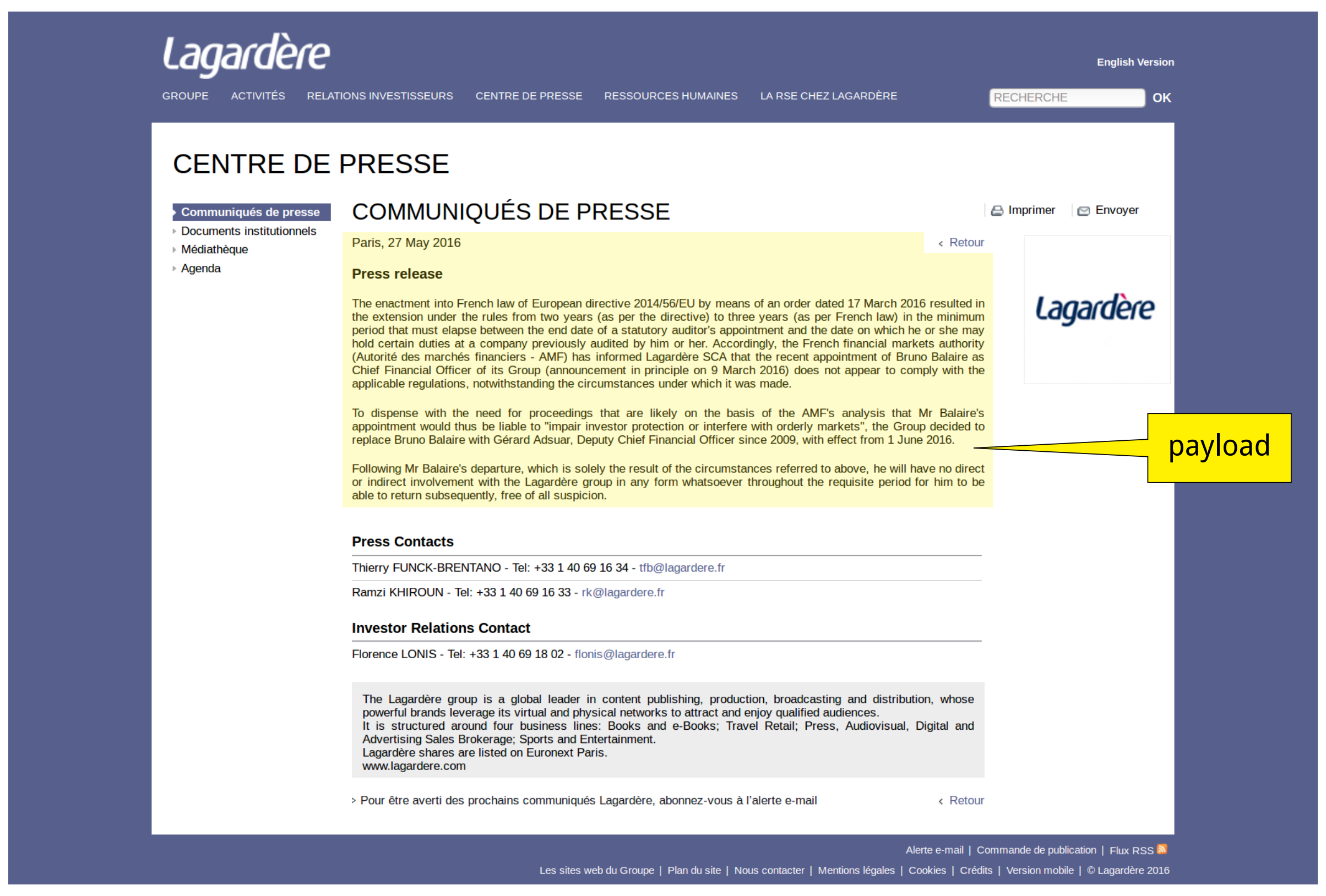
features used			seed data included: strict[a]	1.00[b]	0.99[b]	0.95[b]	0.90[b]	seed data excluded: strict[a]	1.00[b]	0.99[b]	0.95[b]	0.90[b]
cosine (cos)			86.7	93.4	95.4	96.7	97.6	82.5	88.9	91.3	92.9	93.7
"local" cos. (lcos)			86.7	92.8	94.7	95.8	96.9	83.3	88.9	91.4	92.8	93.6
URL similarity (url)			83.6	87.8	88.1	88.2	88.2	83.6	87.8	88.1	88.2	88.2
cos	lcos		87.2	93.7	95.6	96.6	97.5	83.3	89.7	92.1	93.6	94.4
cos		url	90.6	94.7	95.6	96.4	97.1	86.3	90.6	91.4	92.7	93.5
	lcos	url	91.3	95.4	96.3	97.2	97.8	86.8	91.3	92.2	93.4	94.2
cos	lcos	url	92.8	96.7	97.6	98.5	99.1	88.0	92.5	93.4	94.7	95.5
recall on the test data												
cos	lcos	url						87.6	87.6	94.1	95.5	96.0

[a] exact string match with the reference ULR pairs

[b] soft match based on document similarity with different thresholds.

Special Characteristics of Web Pages

- Web pages are a mix of **boilerplate text** and **payload**. **Boilerplate text** (menus, disclaimers, etc.) occurs on most web pages that belong to a particular (sub-)domain. The **payload** is the actual "content" of the page. Boilerplate text is very specific to specific domains but provides little information about the payload.



Lagardère

English Version

GROUPE ACTIVITÉS RELATIONS INVESTISSEURS CENTRE DE PRESSE RESSOURCES HUMAINES LA RSE CHEZ LAGARDÈRE

RECHERCHE OK

CENTRE DE PRESSE

- Communiqués de presse
- Documents institutionnels
- Médiathèque
- Agenda

COMMUNIQUÉS DE PRESSE

Imprimer Envoyer

Paris, 27 May 2016

Retour

Press release

The enactment into French law of European directive 2014/56/EU by means of an order dated 17 March 2016 resulted in the extension under the rules from two years (as per the directive) to three years (as per French law) in the minimum period that must elapse between the end date of a statutory auditor's appointment and the date on which he or she may hold certain duties at a company previously audited by him or her. Accordingly, the French financial markets authority (Autorité des marchés financiers - AMF) has informed Lagardère SCA that the recent appointment of Bruno Balaire as Chief Financial Officer of its Group (announcement in principle on 9 March 2016) does not appear to comply with the applicable regulations, notwithstanding the circumstances under which it was made.

To dispense with the need for proceedings that are likely on the basis of the AMF's analysis that Mr Balaire's appointment would thus be liable to "impair investor protection or interfere with orderly markets", the Group decided to replace Bruno Balaire with Gérard Adsuar, Deputy Chief Financial Officer since 2009, with effect from 1 June 2016.

Following Mr Balaire's departure, which is solely the result of the circumstances referred to above, he will have no direct or indirect involvement with the Lagardère group in any form whatsoever throughout the requisite period for him to be able to return subsequently, free of all suspicion.

payload

Press Contacts

Thierry FUNCK-BRENTANO - Tel: +33 1 40 69 16 34 - tfb@lagardere.fr

Ramzi KHIROUN - Tel: +33 1 40 69 16 33 - rk@lagardere.fr

Investor Relations Contact

Florence LONIS - Tel: +33 1 40 69 18 02 - flonis@lagardere.fr

The Lagardère group is a global leader in content publishing, production, broadcasting and distribution, whose powerful brands leverage its virtual and physical networks to attract and enjoy qualified audiences.
It is structured around four business lines: Books and e-Books; Travel Retail; Press, Audiovisual, Digital and Advertising Sales Brokerage; Sports and Entertainment.
Lagardère shares are listed on Euronext Paris.
www.lagardere.com

Pour être averti des prochains communiqués Lagardère, abonnez-vous à l'alerte e-mail

Retour

Alerte e-mail | Commande de publication | Flux RSS

Les sites web du Groupe | Plan du site | Nous contacter | Mentions légales | Cookies | Crédits | Version mobile | © Lagardère 2016

- Different URLs may lead to the same (or almost the same) page. This leads to problems if they rely only on pairs of URLs for evaluation.
- Web pages are often dynamically generated, sometimes delivering the same payload with different "facades" (e.g., web view, print view, boilerplate in different languages).

http://www.lagardere.com/centre-presse/communiques-de-presse/communiques-de-presse-122.html ...

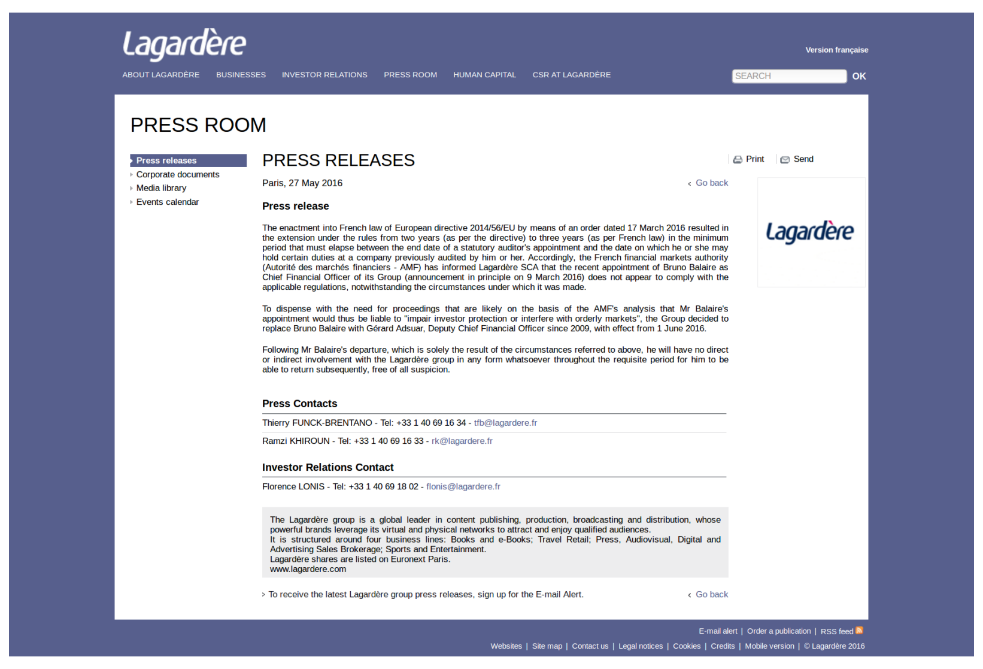
... &idpress=7124

... &idpress=7125

http://www.lagardere.com/press-room/press-releases/press-releases-363.html ...

... &idpress=7124

... &idpress=7125



This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No 645487 (MMT) and 688139 (SUMMA).