Correlation clustering for crosslingual link detection

Jurgen Van Gael, Xiaojin Zhu | Department of Computer Sciences
University of Wisconsin, Madison, Wisconsin, USA

Introduction

Our research solves the crosslingual link detection problem in news processing. This problem calls for identifying news articles in multiple languages that report on the same news event. Our work uses machine translation tools in combination with a recent clustering technique that incorporates cannot-link constraints. Our work handles many topics by applying scaling techniques from optimization. In contrast to existing techniques, our work is flexible in the number of language supported by using machine translation tools. An example crosslingual linked topic is:

Ausschreitungen nach US-Angriffen | Schwere Ausschreitungen mit Toten in Kabul | Ausschreitungen mit 20 Toten in Kabul | ...
Émeutes à Kaboul | Émeutes meurtrières à Kaboul | L'US Army met le feu à Kaboul | L'Afghanistan dans la tourmente | ...
Violentas protestas y saqueos en Afganistán | Afganos evalúan daños tras disturbios anti estadounidenses | ...
Afghanistan: truppe presidiano Kabul dopo disordini anti USA | Torna tranquillità dopo il coprifuoco | Truppe presidiano Kabul | ...
Brake failure caused crash that sparked Kabul riot | Kabul under curfew after deadly riot | Crash spurs deadly Kabul riot | ...
Protestos em Afeganistão | Acidente de trânsito gera caos em Cabul | Vaga de violência na capital afegã | ...
爆反美怒潮阿富汗宵禁 | 新华社记者采访遭围攻(组图) | 美军公布引发喀布尔骚乱的车祸原因 | 美军车肇事引喀布尔骚乱,中国商店和记者受牵连 | ...
이라크전 희생 민폰인수 2차대전 추월 | 미 CBS 위제전 이라크서 2명 사망, 1명 중태 | CBS 카메라맨, 녹음기사, 바그다드 폭발로 사망 | ...

System Overview

1. Data Collection:
   - Crawl Google News

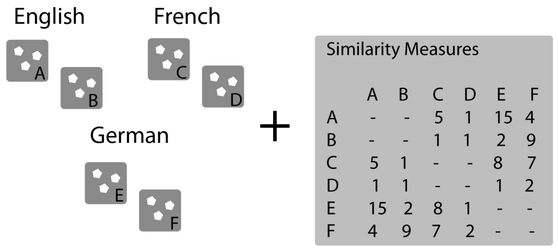
     L'Amérique rend hommage à Gerald Ford
     Washington se recueille sur la dépouille de l`ancien président
     L'hommage des Etats-Unis à Gerald Ford

   - Translate articles to english

     America pays homage to Gerald Ford
     Washington collects on the skin of L `former president the
     homage of the United States to Gerald Ford

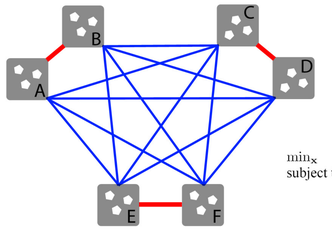
   - Prepare TF.IDF vector
2. Setup correlation clustering (see right)
3. Solve linear program using chunking (see right)
4. Round off linear program solution (see right)

Correlation Clustering with chunking



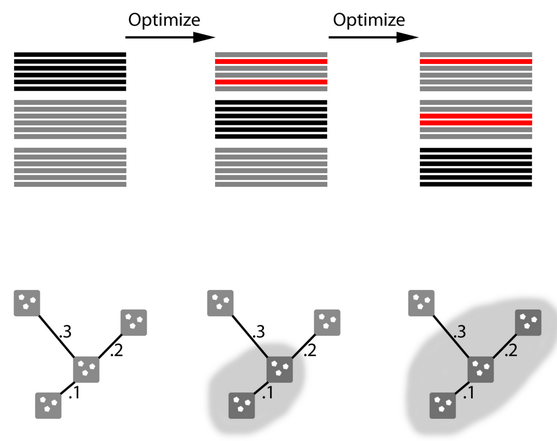
Similarity Measures

	A	B	C	D	E	F
A	-	-	5	1	15	4
B	-	-	1	1	2	9
C	5	1	-	-	8	7
D	1	1	-	-	1	2
E	15	2	8	1	-	-
F	4	9	7	2	-	-



Positive Weight
Negative Weight

$$\begin{aligned} \min_{\mathbf{x}} \quad & \textstyle\sum_{e \in E} w_e x_e \\ \text{subject to} \quad & x_e \in \{0,1\}, && \forall e \in E \\ & x_{uv} + x_{vt} \geq x_{ut}, && \forall uv, vt, ut \in E \\ & x_{uv} = x_{vu}, && \forall u, v \in V \end{aligned}$$



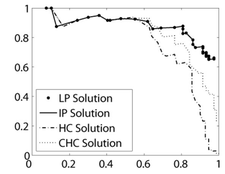
Phase I - Data Collection
We gather news article titles and group them per language by article topic. All article titles are translated to English and represented as a TF.IDF vector. Similarity measures are computed by calculating the angle between TF.IDF vectors.

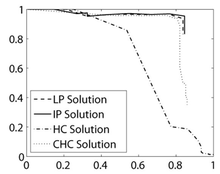
Phase II - Correlation Clustering
We setup the linear program as follows. All article groups represent nodes in a graph. We introduce a binary variable for every weighted edge in the graph. The magnitute of the edge weights represents how strong we believe that the article groups cover the same topic. We put a large negative weight on edges between article groups in the same language. The binary variables indicate whether two groups cover the same topic (1) or not (0). The only integer programming constraint is a triangle inequality constraint making sure the solution encodes a valid clustering. Finally, the objective function minimizes the positive weight *across* clusters.

Phase III - Chunking
We relax the integer program to a linear program and solve it using chunking. We split the constraints in different chunks and consider only one constraint chunk (black). We optimize and remember the active constraints (red) after which we take the next chunk into consideration. We iterate until convergence.

Phase IV - Rounding
We round the linear programming solution to binary variables which encode a clustering. The rounding is calculated by growing balls around the article groups according to the edge variables

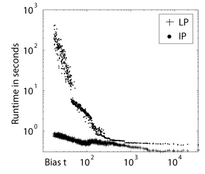
Results





Left: average precision recall on 4 datasets linking 60 French, German and English news articles. Right: precision recall curve on a dataset linking 160 English, German, Italian, French, Portuguese, Spanish, Korean and Chinese. HC is standard hierarchical clustering, CHC is constrainined hierarchical clustering.

This plot compares the runtime of solving the exact integer program compared to the performance of the linear program with chunking.



Discussion

Our work shows good results on the crosslingual link detection problem for datasets with up to 500 article topics despite mediovre machine translation tools. We believe performance could be improved by including extracted features such as people and geographical names into our article representation rather than using more sophisticated machine translation tools. Finally, our work shows the strength and weaknesses of the correlation clustering framework for the first time.

Contact

Contact Jurgen Van Gael at jvangael@cs.wisc.edu for more information. Datasets available at:
http://www.cs.wisc.edu/~jvangael/newsdata/