New Methods for Graph Matching with Applications in Digital Libraries

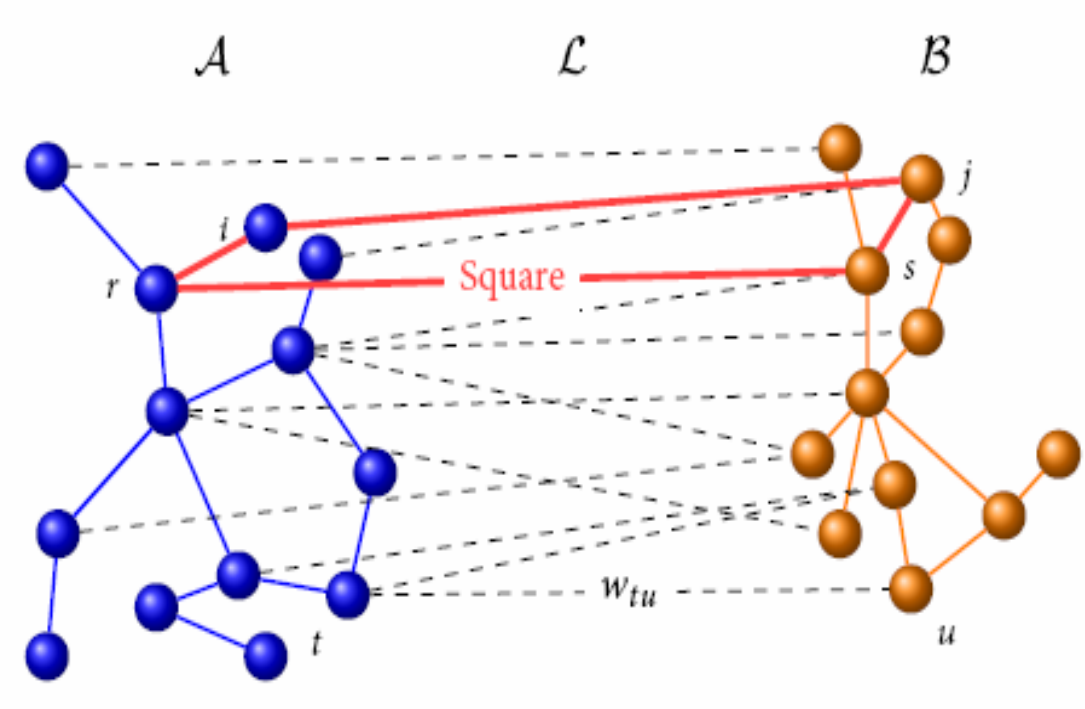
Mohsen Bayati, Margot Gerritsen,
David Gleich, Amin Saberi, Ying Wang

Problem Formulation

Input: Two graphs A and B. A set of potential matches L.

Output: A matching.

Objective:
Maximize overlap (number of squares).



Applications:

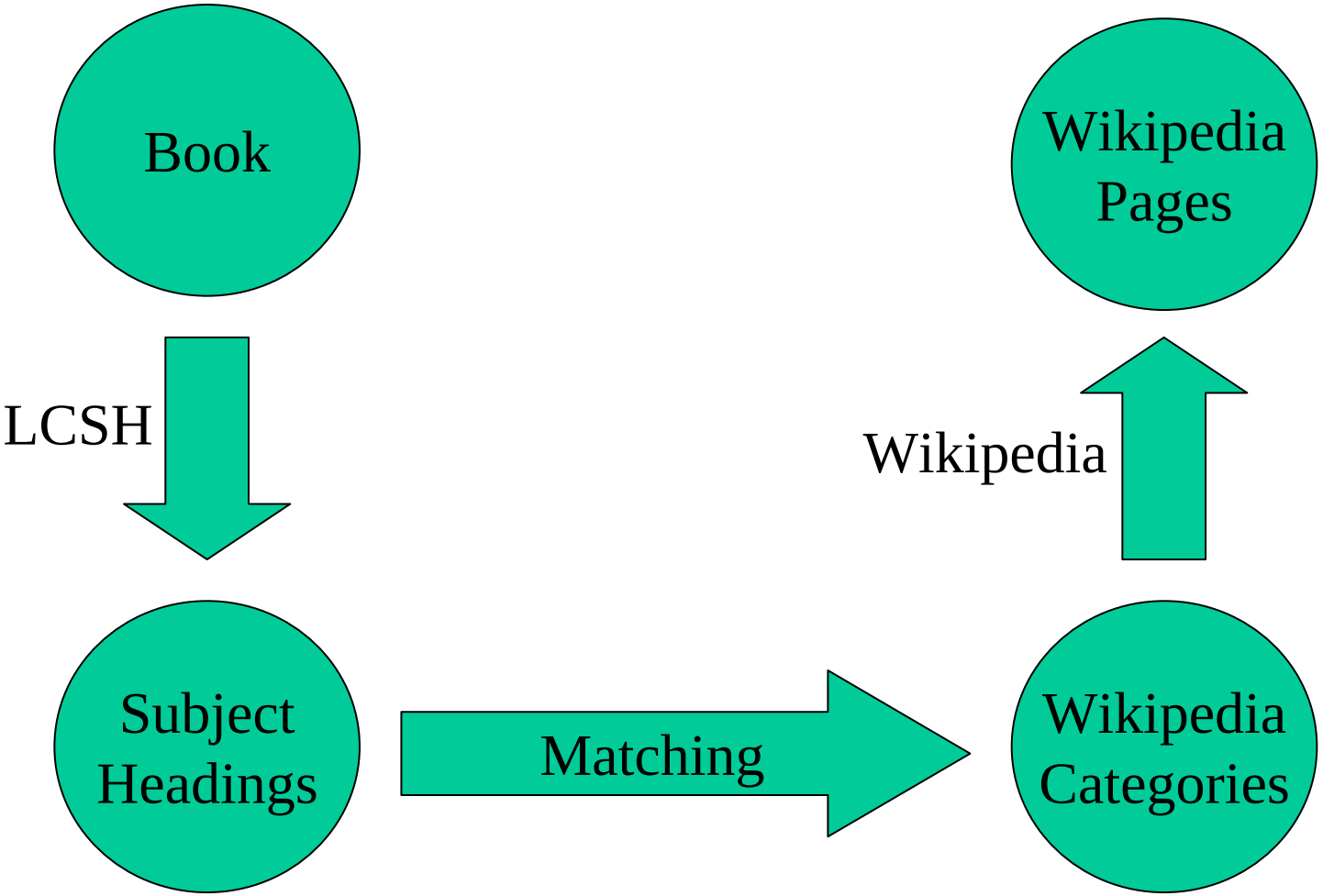
Matching LCSH to Wikipedia

Multilingual WordNet

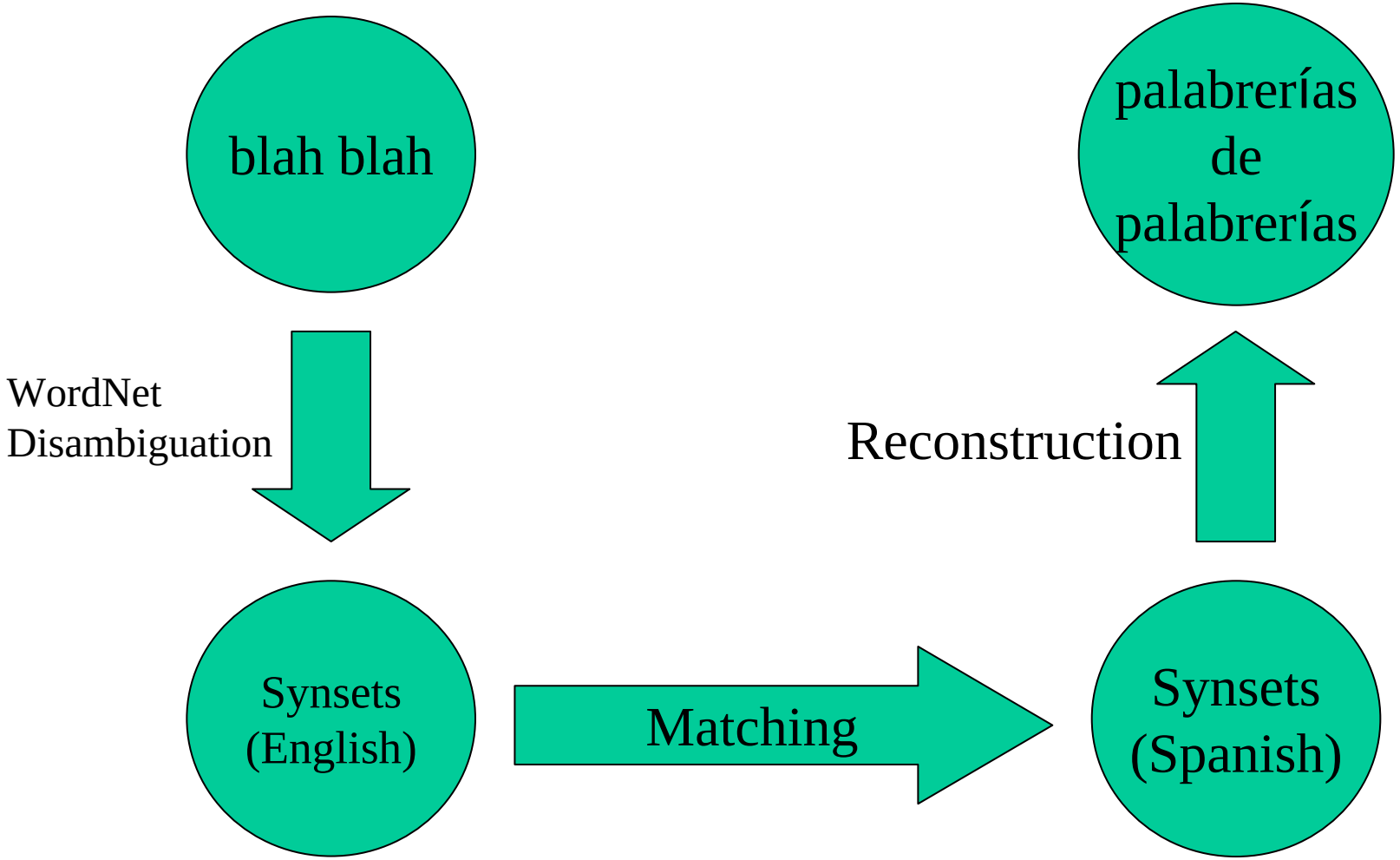
Why Graph Matching?

- Finding implicit relationships between large datasets
- Compute similarity of different networks
- The matching is a portal between datasets

Graph Matching Application: Finding Wikipedia Pages Related to a Book



Graph Matching Application: Multilingual Translation



Graph Matching Applications

- *Ontology matching (Matching LCSH to Wikipedia)*
- *Multilingual WordNet*
- Database schema matching
- Pattern recognition
- PPI (protein-protein interaction) network alignment
- Chemical structure matching

Graph Matching: Algorithmic Consideration

- NP-hard (very hard to compute the optimal solution)
- APX-hard (also hard to approximate)
- We use BP (belief propagation) to get a good solution, and use LP (linear program) to get a good upper bound.

Existing Methods

- Similarity flooding (Melnik et al, 2002)
- IsoRank (Singh et al, 2007)
- Lagrangean decomposition (Klau, 2009)

Our Algorithmic Contribution

- We consider the sparse case of the problem, which enables the possibility of approaching large-scale data sets.
- We propose a new algorithm based on BP (belief propagation) that outperforms existing algorithms by around 40% in most cases.
- We also studied existing algorithms and compared their performances on various data sets.

LCSH to Wikipedia Categories: Similarities and Differences

	LCSH	Wikipedia Category
Similarities	subject heading	category page
	narrower term	subcategory
	broader term	super category
Differences	catalog experts	“non-experts”
	long development period	short development period
	centralized	distributed

LCSH to Wikipedia Categories: Problem Scale

Problem		Graph	Vertices	Edges
Small	$\mathcal{A}$	LCSH	1,919	3,130
	$\mathcal{B}$	WC	2,000	7,811
	$\mathcal{L}$		3,919	16,952
Full	$\mathcal{A}$	LCSH	297,266	248,232
	$\mathcal{B}$	WC	226,221	422,503
	$\mathcal{L}$		523,487	5,233,829

LCSH to Wikipedia Categories: Computational Results

α	β	Algorithm	Objective	Overlap	Weight	Cardinality
1	0	MWM	60119.8	2346	60119.8	106294
0	1	LP	28660.4	——–	——–	——–
		LP-x	11025	11025	46265.2	96578
		LP-Y	14176	14176	46353.5	96304
		LP-zY	14253	14253	46327.4	96287
		BP	13727	13727	37387.7	76136
		BP-Y	12721	12721	46268.1	96262
		BP-zY	13392	13392	46199.1	96141
		ISORANK	9539	9539	55697.3	103146
		ISORANK-S	11834	11834	46558.3	96309
		MWM	2346	2346	60119.8	106294

The current best solution we get from BP has an overlap above 15,000 and it outperforms Isorank by about 50%.

LCSH to Wikipedia Categories: Conclusions

- Wikipedia categories has a certain level of similarity to LCSH. (about 10% overlap)
- However, they are also different in structure and a good matching can not be found.
- Question 1: What makes them different?
- Question 2: Can we identify regions where they are similar?

Multilingual WordNet Alignment: What is WordNet?

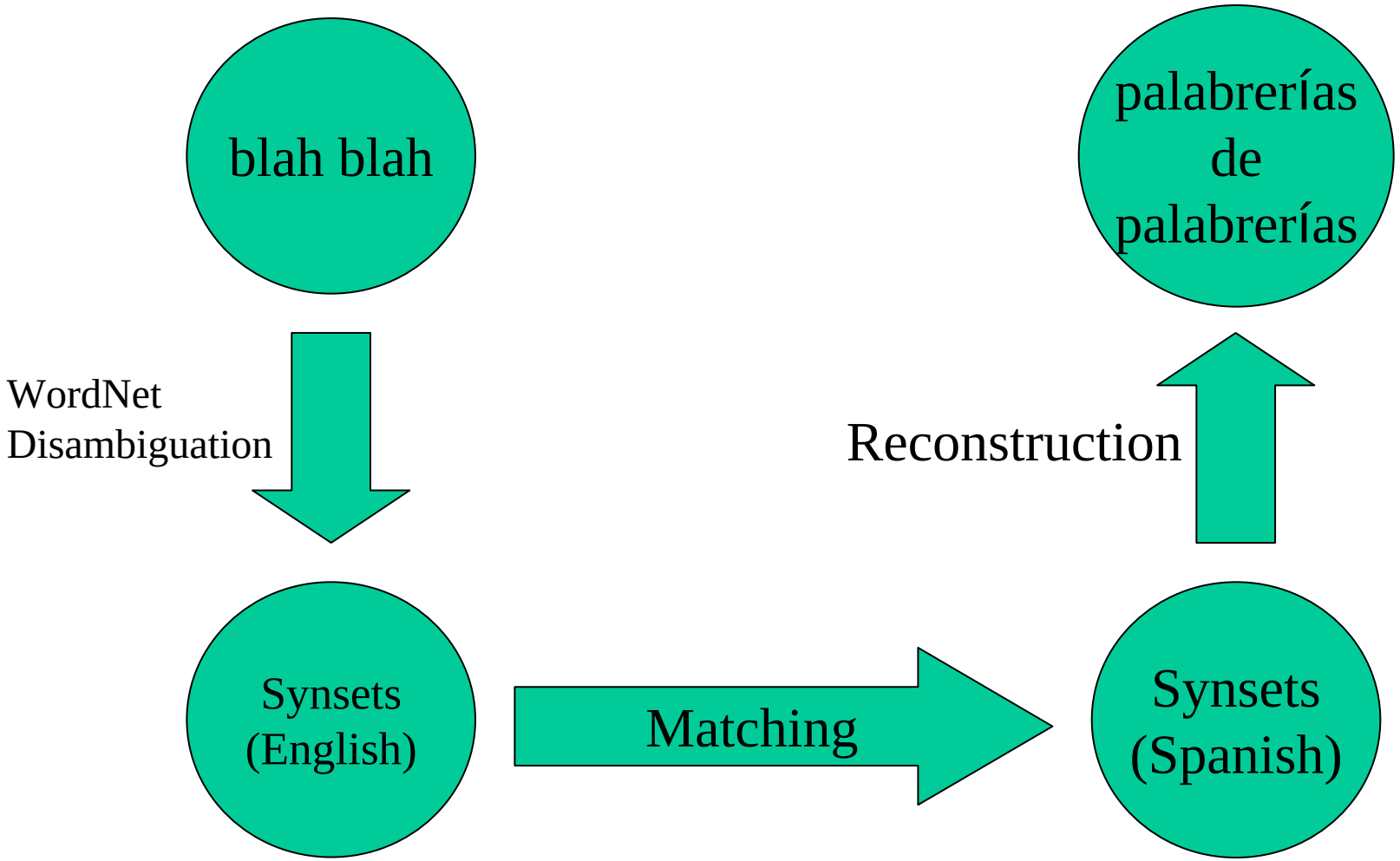
- Synset: A group of words representing the same meaning (cognitive synonyms).
  - S: (v) **help**, aid (improve the condition of) *"These pills will help the patient"*
  - S: (v) **help**, facilitate (be of use) *"This will help to prevent accidents"*
- WordNet divides words into synsets, and builds connections between synsets.
- *WordNet captures the subtleties of the meanings of words.*

Multilingual WordNet Alignment: Our Goal

- Compute similarities of word nets of different languages by graph matching.

- Build a multilingual word net based on the calculated matching.

Multilingual Translation Revisited



Multilingual WordNet Alignment: Experiment Setup

- Potential matches: Two words (one in English and one in Spanish) can match if they appear in the same verse of the Bible.

- A (English): 56697 nodes, 57548 edges.
- B (Spanish): 5536 nodes, 5419 edges.
- L (potential matches): 138302 edges.

Multilingual WordNet Alignment: Results

- Matching size: 1493
- BP overlap: 420
- IsoRank overlap: 300

- Overlap is small!

Multilingual WordNet Alignment: A Sample Matching

bear	---	soportar
cover	---	cubrir
off	---	nube
mark	---	marca
pass	---	alto
passing	---	amor
dark	---	oscuro
get	---	inteligencia
fall	---	ruina
birth	---	nacimiento

Multilingual WordNet Alignment: How can we improve it?

- Matching on synset level instead of word level!
- Better word net database.
- Better way of finding potential matches.

Summary

- We proposed an efficient, scalable algorithm for the classic graph matching problem, and our algorithm outperforms all existing algorithms considerably (about 40%).
- The algorithm has lots of applications in various fields.

Future directions

- Find good approximation algorithm.
- How can we evaluate the result (accuracy)?
- Evaluate the algorithm on more data sets.