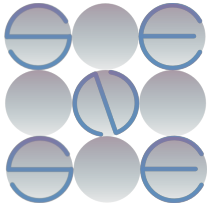


UNIBA-SENSE @ CLEF 2009: Robust WSD task

Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro

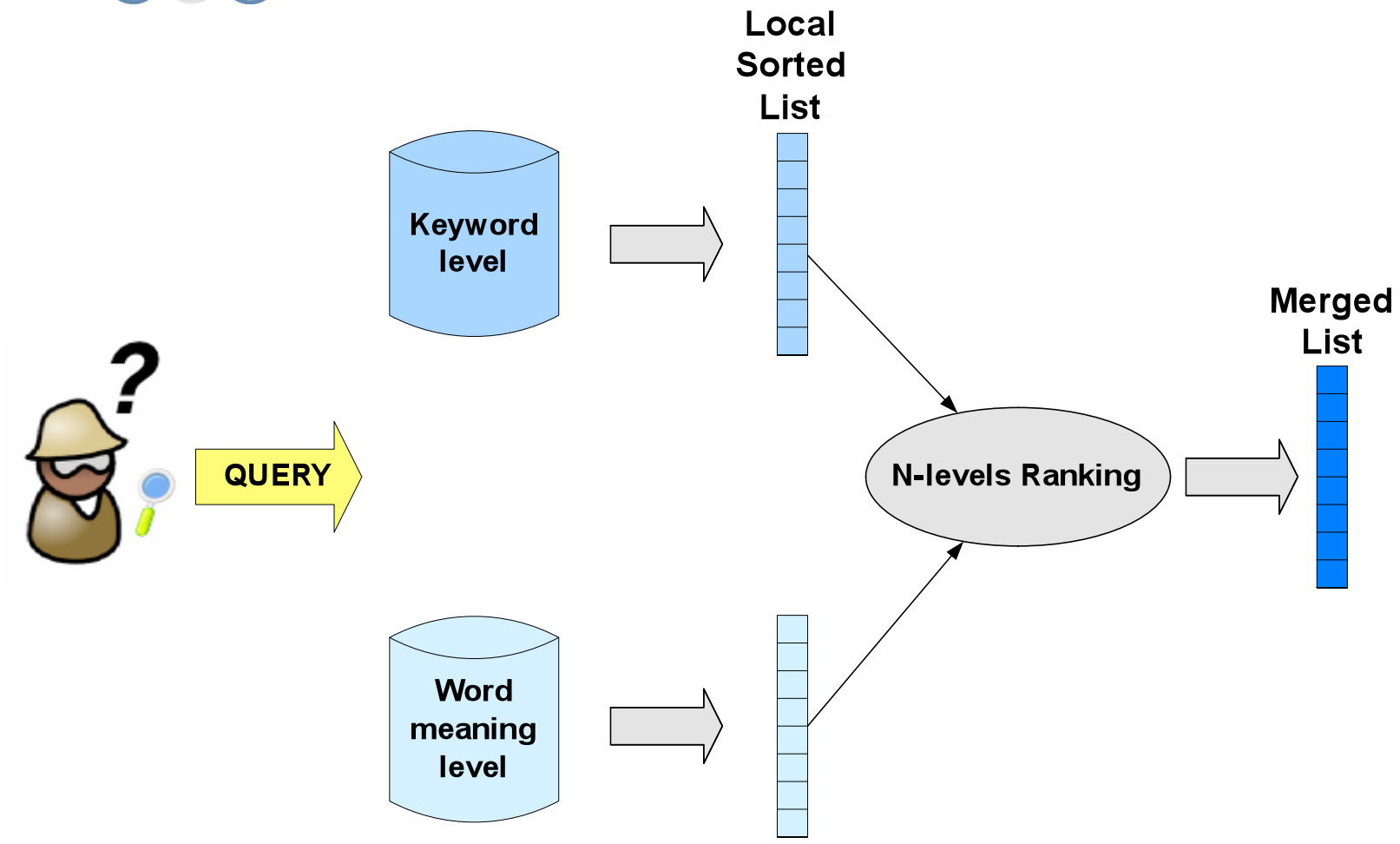
Dept. of Computer Science
University of Bari “Aldo Moro” (ITALY)
{basilepp,acaputo,semeraro}@di.uniba.it

Ad Hoc Robust-WSD@CLEF

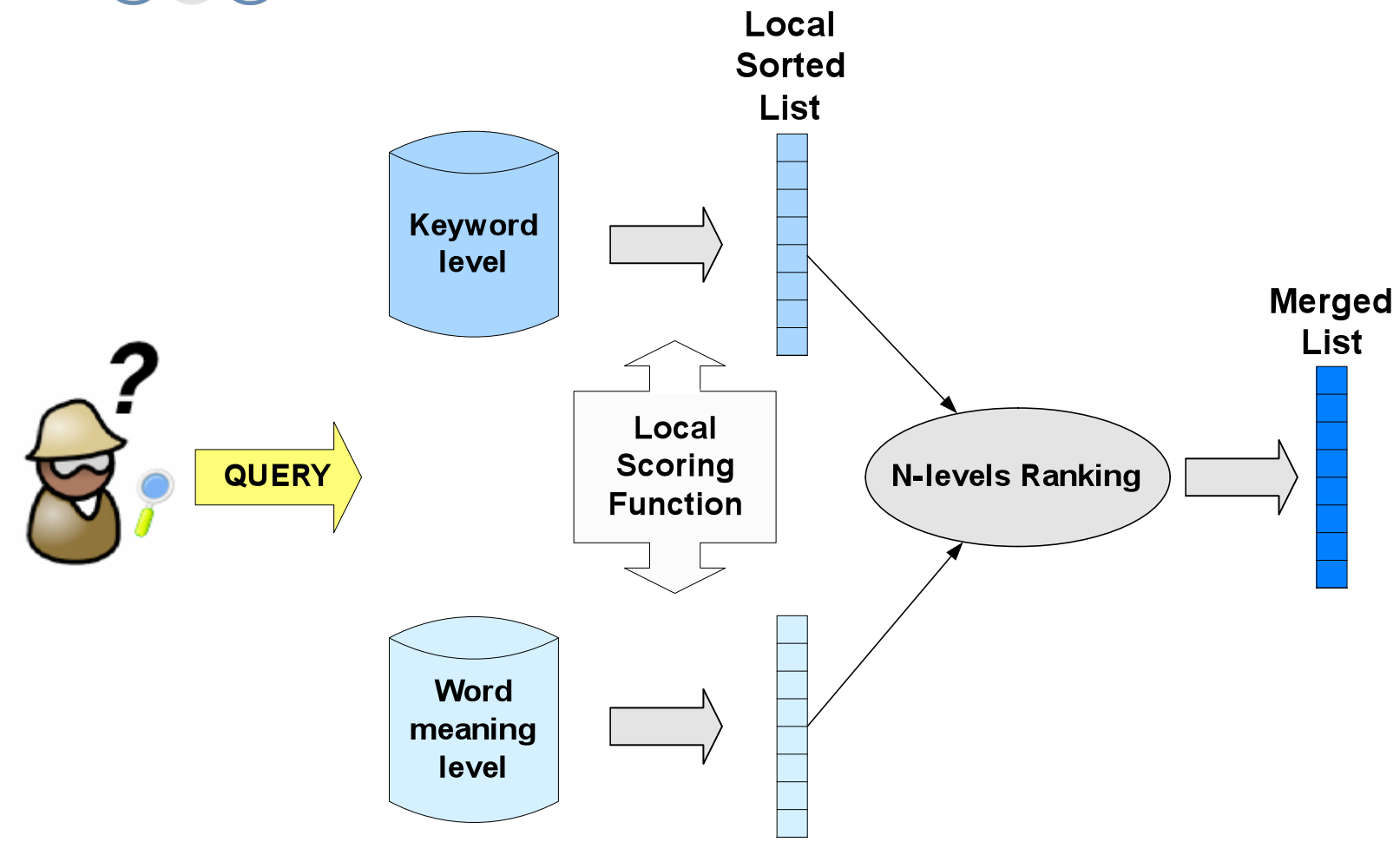
Outline

- SENSE
- Indexing
- Searching
  - Strategy
  - Cross-language
  - Pseudo Relevance Feedback
- Results and Conclusion

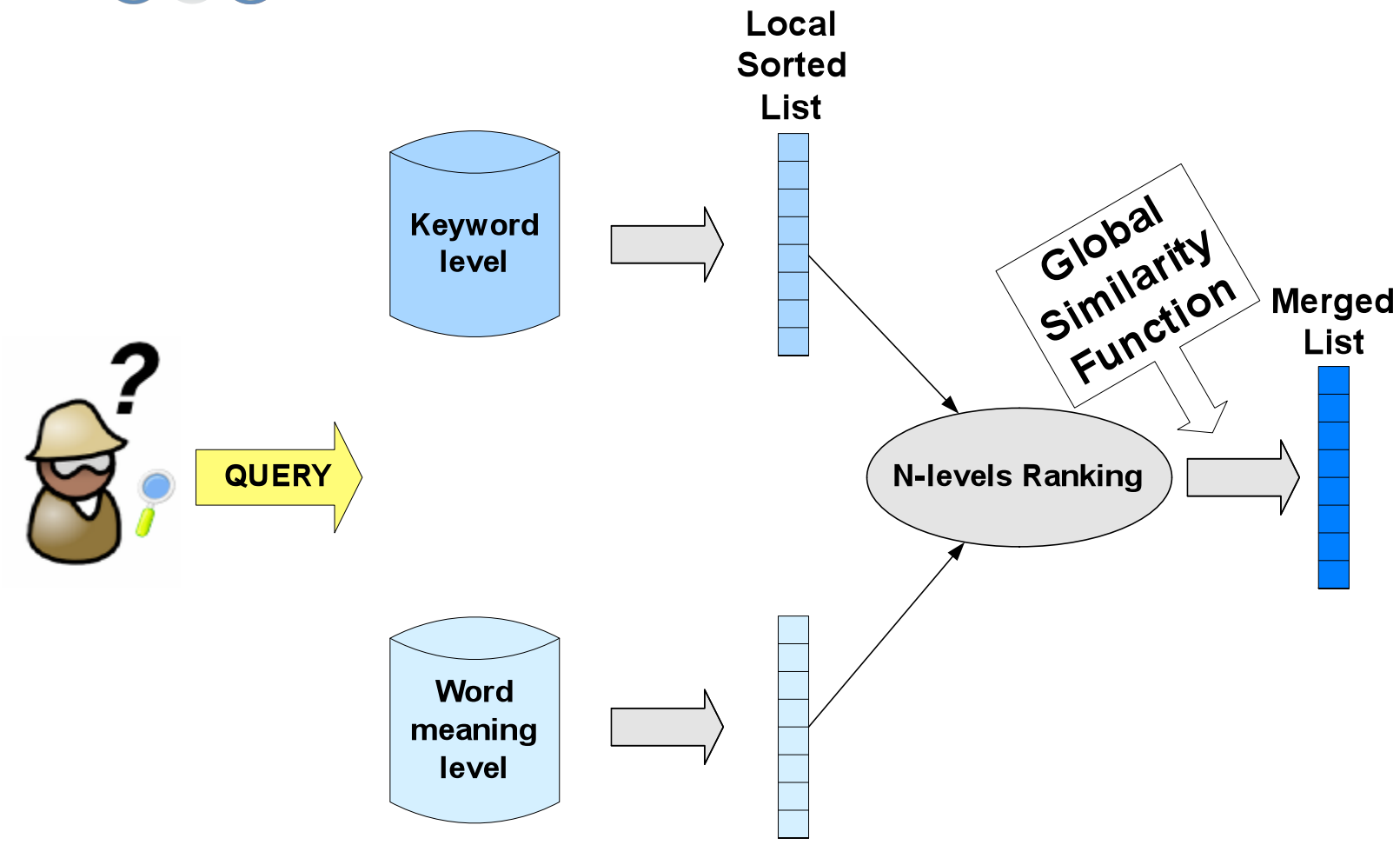
SENSE Architecture



SENSE Architecture



SENSE Architecture



Indexing

- Keyword level
  - Word stemming
  - Stop word elimination
  - Text normalization
    (not alphanumeric characters vs. “_”)
- Word Meaning level
  - Only the synset with the highest score
- Two fields for each document
  - HEADLINE
  - TEXT

Searching

- Local Similarity Function
  - Okapi BM25[1]
    - Keyword
    - Word meaning level
  - Reflects the multi-field representation of documents
    - Headline
    - Text

[1]S. Robertson, H. Zaragoza and M. Taylor. Simple BM25 extension to multiple weighted fields. CIKM '04.

Searching: strategy

- Keyword level
  - Different strategies involving Title (T), Description (D) and Narrative (N) fields of the topic
  - Query expansion by Local Context Analysis
- Word Meaning level
  - Similar to the keyword level
  - Exploited only the synset with highest score
- Levels aggregation: merges result lists using the Global Ranking Function
    - Normalization: Z-score
    - Aggregation: CombSUM

Searching: Cross-language

- Keyword level
  - Automatic translation by Google Translator API
    - Spanish -> English
- Word Meaning level
  - Only the synset with the highest score
    - Always first synset in Spanish WordNet

Searching: Pseudo Relevance Feedback

- Query expansion with Local Context Analysis[1](LCA)
  - Querying each level
    - Get the *n* top ranked documents
    - Extract all features from the documents
    - Adding the top *k* ranked features to the query
    - Re-querying each level

[1]J. Xu and B. Croft. Improving the effectiveness of information retrieval with local context analysis. ACM Trans. Inf. Syst. 2000.

Searching: LCA

$$codegree(f, q_i) = \frac{\log_{10}(co(f, q_i) * idf(f))}{\log_{10}(n)}$$

$$co(f, q_i) = \sum_{d \in S} tf(f, d) * tf(q_i, d)$$

$$idf(f) = \min(1.0, \frac{\log_{10}(N / N_f)}{5.0})$$

- *codegree*: degree of co-occurrence between feature *f* and query q_i
- *Idf(f)*: frequency of *f* in the collection
- *n*: number of documents in the top-ranked set
- *tf*: term frequency in document *d*
- *S*: set of top ranked documents
- *N*: collection size
- N_f: number of documents containing *f*

Searching: LCA

$$lca(f,q) = \prod_{q_i \in q} (\delta + codegree(f,q_i))^{idf(q_i)}$$

- Features, from the first (*n*=10) documents, ranked according to *lca*
- *δ* smoothing factor set up to 0.1
- Added the first *k* (*k*=10) features to the original query (both for keyword and word meaning)
  - *lca* as features weight

Searching: parameters

	Field	k_1	N	avl_c	b_c	$boost_c$
BM25	$HEADLINE_{keyword}$	3.25	166,726	7.96	0.70	2.00
	$TEXT_{keyword}$	3.25	166,726	295.05	0.70	1.00
	$HEADLINE_{synset}$	3.50	166,726	5.94	0.70	2.00
	$TEXT_{synset}$	3.50	166,726	230.54	0.70	1.00

LCA	δ	k	n
	0.1	10	10

- Aggregation weight for the two levels
  - Keyword=0.8 ⋀ Word Meaning=0.2
  - Keyword=0.9 ⋀ Word Meaning=0.1

Results: System setup

				RUN
MONO LINGUAL	NO WSD	ONLY KEYWORD	NO LCA	unibaKTD unibaKTDN
			LCA	unibaKRF
	WSD	ONLY SYNSET	No LCA	unibaWsdTD unibaWsdTDN
		N-LEVELS	No LCA	unibaWsdNL0802 unibaWsdNL0901
			LCA	unibaKeySynRF
BILINGUAL	No WSD	ONLY KEYWORD	No LCA	unibaCrossTD unibaCrossTDN
			LCA	unibaCrossKeyRF
	WSD	ONLY SYNSET	No LCA	unibaCrossWsdTD unibaCrossWsdTDN
		N-LEVELS	No LCA	unibaCrossWsdNL0802 unibaCrossWsdNL0901
			LCA	unibaCrossWsdKeySynRF

Results: Monolingual

RUN	MAP	GMAP
unibaKTD	.3963	.1684
unibaKTDN	.4150	.1744
unibaKRF	.4250	.1793
unibaWsdTD	.2930	.1010
unibaWsdTDN	.3238	.1234
unibaWsdNL0802	.4218	.1893
unibaWsdNL0901	.4222	.1864
unibaKeySynRF	**.4346**	**.1960**

Results: Monolingual

RUN	MAP	ΔMAP	GMAP	ΔGMAP
unibaKTD	.3963		.1684	
unibaKTDN	.4150		.1744	
unibaKRF	.4250		.1793	
unibaWsdTD	.2930		.1010	
unibaWsdTDN	.3238		.1234	
unibaWsdNL0802	.4218		.1893	
unibaWsdNL0901	.4222		.1864	
unibaKeySynRF	**.4346**	+2.26	**.1960**	+9.31

Baseline

Results: Bilingual

RUN	MAP	GMAP
unibaCrossTD	.3414	.1131
unibaCrossTDN	.3731	.1281
unibaCrossKeyRF	**.3809**	.1311
unibaCrossWsdTD	.0925	.0024
unibaCrossWsdTDN	.0960	.0050
unibaCrossWsdNL0802	.3675	.1349
unibaCrossWsdNL0901	.3731	.1339
unibaCrossWsdKeySynRF	.3753	**.1382**

Results: Bilingual

RUN	MAP	ΔMAP	GMAP	ΔGMAP
unibaCrossTD	.3414		.1131	
unibaCrossTDN	.3731		.1281	
unibaCrossKeyRF	**.3809**		.1311	
unibaCrossWsdTD	.0925		.0024	
unibaCrossWsdTDN	.0960		.0050	
unibaCrossWsdNL0802	.3675		.1349	
unibaCrossWsdNL0901	.3731		.1339	
unibaCrossWsdKeySynRF	.3753	-1.47	**.1382**	+5.41

Baseline

Conclusion

- Combination of keyword and word meaning improves IR performance
  - Improvement in MAP and GMAP in monolingual task
  - Only GMAP in bilingual task
- LCA is effective
  - Even on word meaning

Thanks for your attention