Automatic Domain Assignment for Word Sense Alignment

Tommaso Caselli[1,2], Carlo Strapparava[3]
Trento RISE[1], VUA[2], Fondazione Bruno Kessler[3]
t.caselli@gmail.com,strappa@fbk.eu

Lexical Knowledge for NLP

Lexical Knowledge is available in different forms:

- unstructured terminologies
- computational lexicon
- ontologies

Issues:

- costs (money and time)
- scattered information
- need for semantic interoperability and reusability

Word Sense Alignment

Word Sense Alignment (WSA): a solution to semantic interoperability.

WSA: a list of pairs of senses from two or more() lexical-semantic resources. A pair of aligned senses denotes the same meaning.

day = *amount of hours of work done in one day* [SC Lexicon]

day = *the recurring hours established by contract or usage for work* [MWN]

Target Resources for Alignment

MultiWordNet (MWN; Pianta et al., 2002):
- Italian version of Princeton WordNet (Fellbaum, 1998)
- Obtained through the "expand model" (Vossen, 1996)
- Aligned to WN 1.6
- Reduced number of sense descriptions in Italian: only 8,21% sense descriptions are in Italian, the remaining are in English and imported from WN 1.6
- Gaps in senses

Senso Comune Lexicon (SCL; Vetere et al., 2011)
- Machine readable dictionary obtained from a paper-based reference lexicographic dictionary (De Mauro GRADIT)
- no taxonomy of senses
- absence of domain or category labels associated with senses
- no distinction between core senses and subsenses for polysemous entries

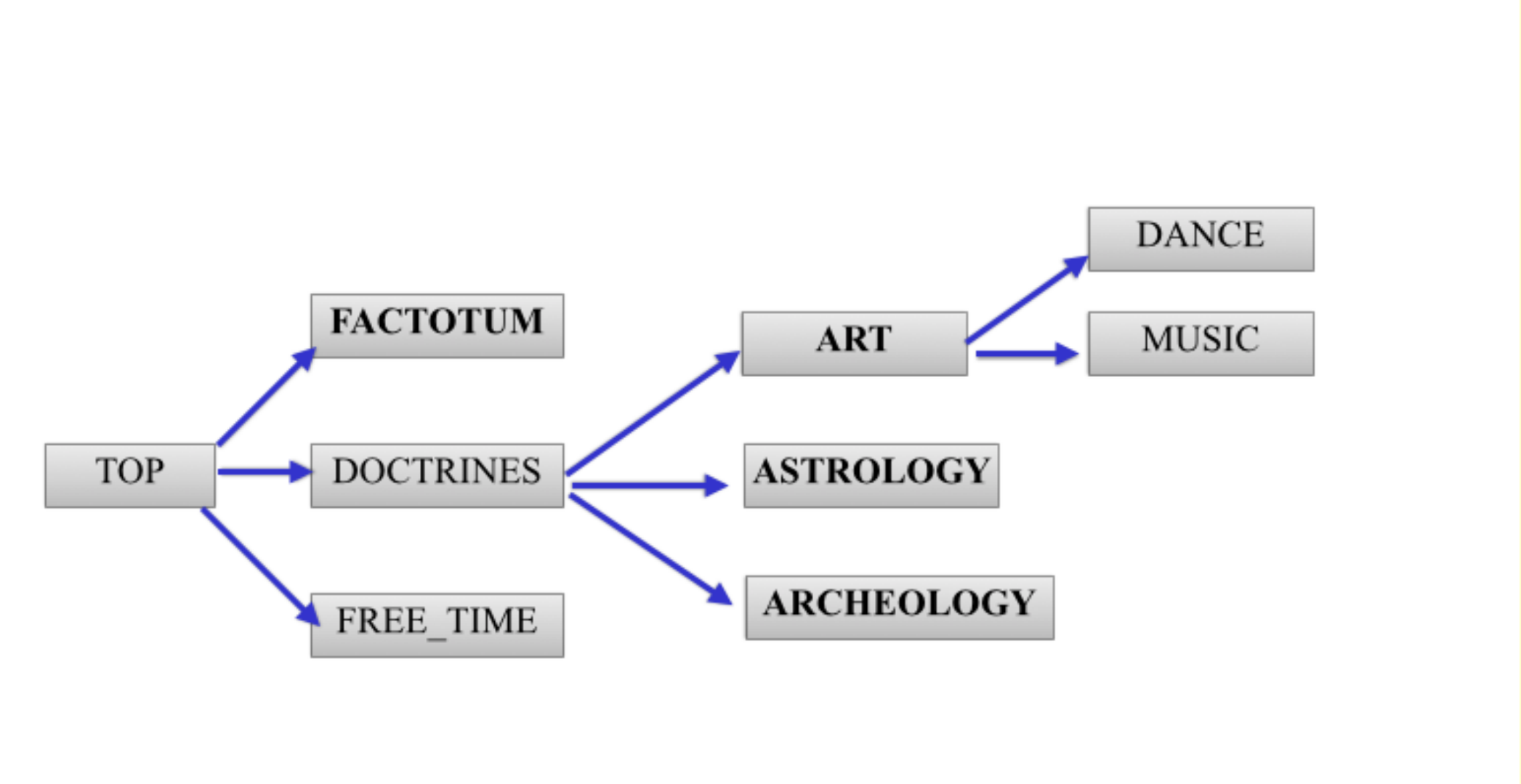
Domain Information and WSA

Domain information associated to word senses is an important feature for improving the quality of aligned entries (Navigli, 2006; Toral et al., 2009; Navigli and Ponzetto, 2012).

WordNet Domains (WN Domains; Magnini and Cavaglia, 2002: Bentivogli et al., 2004) has been selected as a reference domain repository for automatically assigning domain labels to SCL entries.

WN Domains is a lexical resource created in a semi-automatic way by augmenting WordNet synsets with domain labels.

166 hierarchically organised labels expressing a subject field label (SPORT, MEDICINE etc.). The label FACTOTUM is used for those synsets which can appear in almost every domain. Following (Magnini et al., 2001) only 45 domain labels have been used (i.e. normalised domains).



Automatic Assignment of Domain Labels to SCL glosses

Issue #1: can we automatically assign domain information from WN Domains to the sense descriptions of the Senso Comune Lexicon?

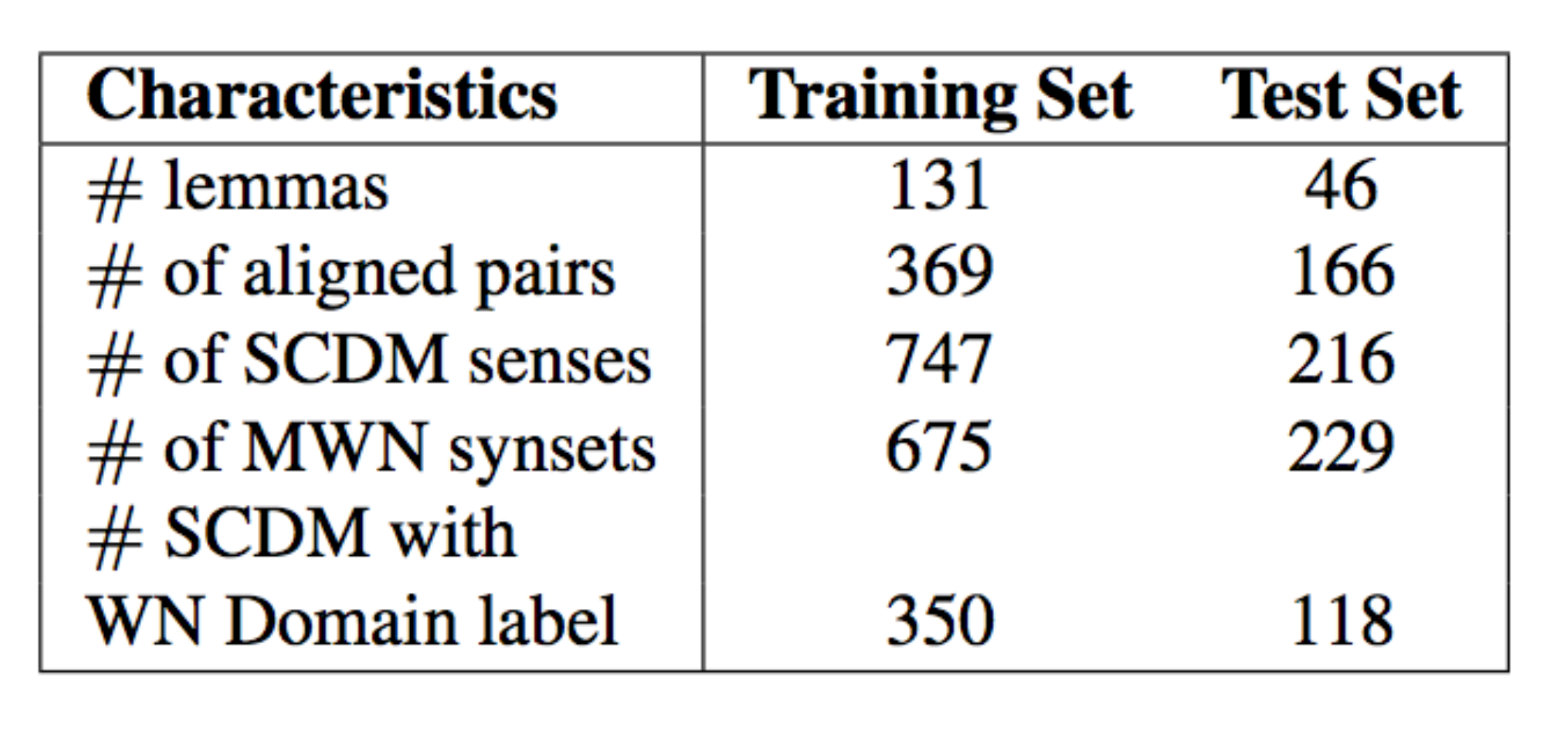
Working Hypothesis: "One Domain per Discourse" hypothesis can be applied to short texts such as sense descriptions in a lexicographic dictionary. This will provide entries from the Senso Comune Lexicon with the same set of 45 normalised domain labels as in MWN.

Approach: (binary) classifier + (post-processing) rules

Target items: nominal entries

Classifier and Feature Selection

Binary classifier:
- FACTOTUM
- OTHER (SPORT, MEDICINE, ASTROLOGY …)



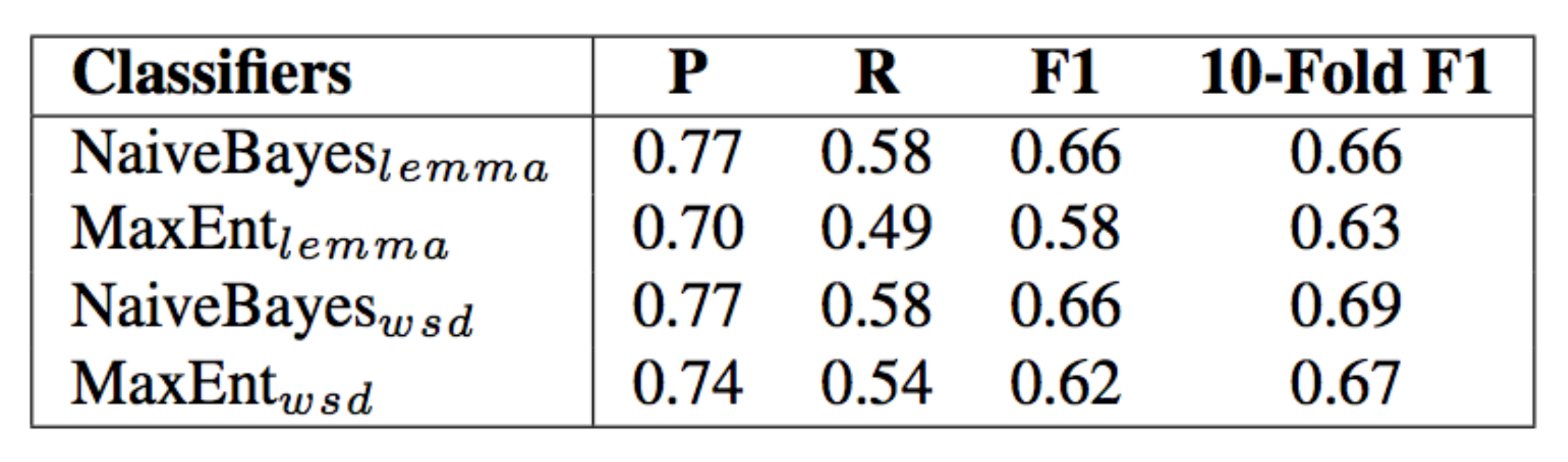
Characteristics	Training Set	Test Set
# lemmas	131	46
# of aligned pairs	369	166
# of SCDM senses	747	216
# of MWN synsets	675	229
# SCDM with WN Domain label	350	118

Assignment of the WN Domain labels to SCL entries by means of manual sense alignment. Test set from Caselli et al., 2014.

Each SCL entry was represented by a two-dimensional feature vector (GENERIC:*val* SPECIFIC:*val*):

- *lemma label*: for each lemma in the sense description, we associated all normalised domains from MWN. Feature values: frequency counts
- *word sense label*: Word Sense Disambiguation (WSD; UKB package; Agirre et al., 2014) of the sense description, retain as good the highest ranked sense in MWN and assign the corresponding domain label. Feature values: frequency counts

Two classifiers: NaiveBayes *vs.* Maximum Entropy



Classifiers	P	R	F1	10-Fold F1
NaiveBayes$_{lemma}$	0.77	0.58	0.66	0.66
MaxEnt$_{lemma}$	0.70	0.49	0.58	0.63
NaiveBayes$_{wsd}$	0.77	0.58	0.66	0.69
MaxEnt$_{wsd}$	0.74	0.54	0.62	0.67

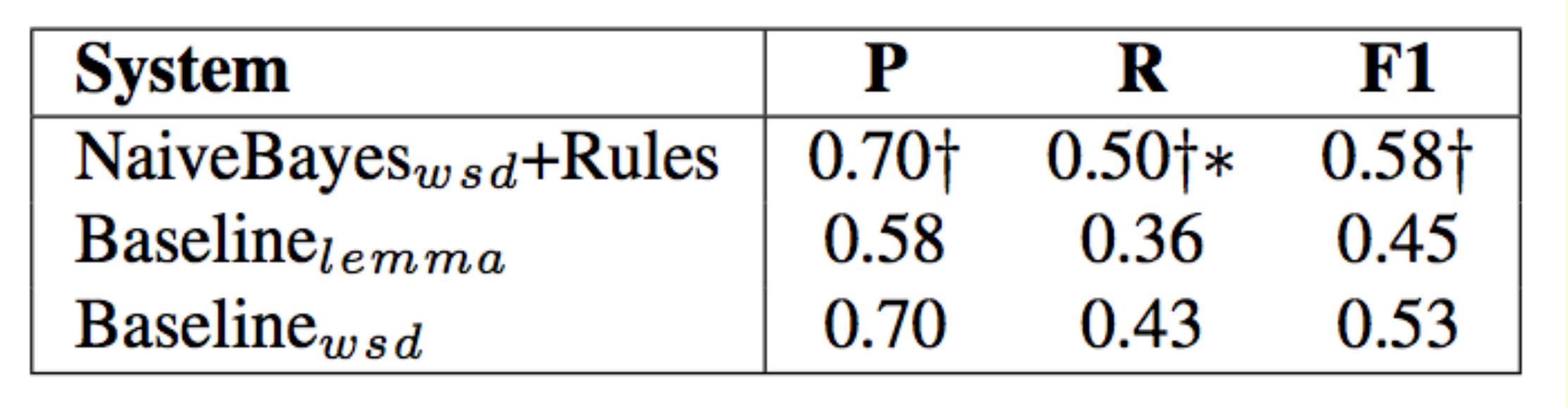
NaiveBayes outperforms Maximum Entropy

Positive role of WSD

Post-processing Rules

Rules apply only to entries classified as OTHER by the NaiveBayes$_{wsd}$. Rules assigne fine-grained domain values (i.e. one or more of the 45 normalised domain labels).

- WSD on sense description classified as OTHER and extraction of the corresponding WN Domain label
- Frequency counts on the domain labels: assign as correct the most frequent domain label
- If frequency score of WN Domain labels equals 1, assign FACTOTUM; if the score is higher than 1, retain all domain labels as good



System	P	R	F1
NaiveBayes$_{wsd}$+Rules	0.70†	0.50†∗	0.58†
Baseline$_{lemma}$	0.58	0.36	0.45
Baseline$_{wsd}$	0.70	0.43	0.53

- Baseline$_{lemma}$: assign one of the 44 normalised domain label by taking into account all domain labels associate to each lemma
- Baseline$_{wsd}$: assign one of the 44 normalised domain label by taking into account domain labels of WSD lemmas

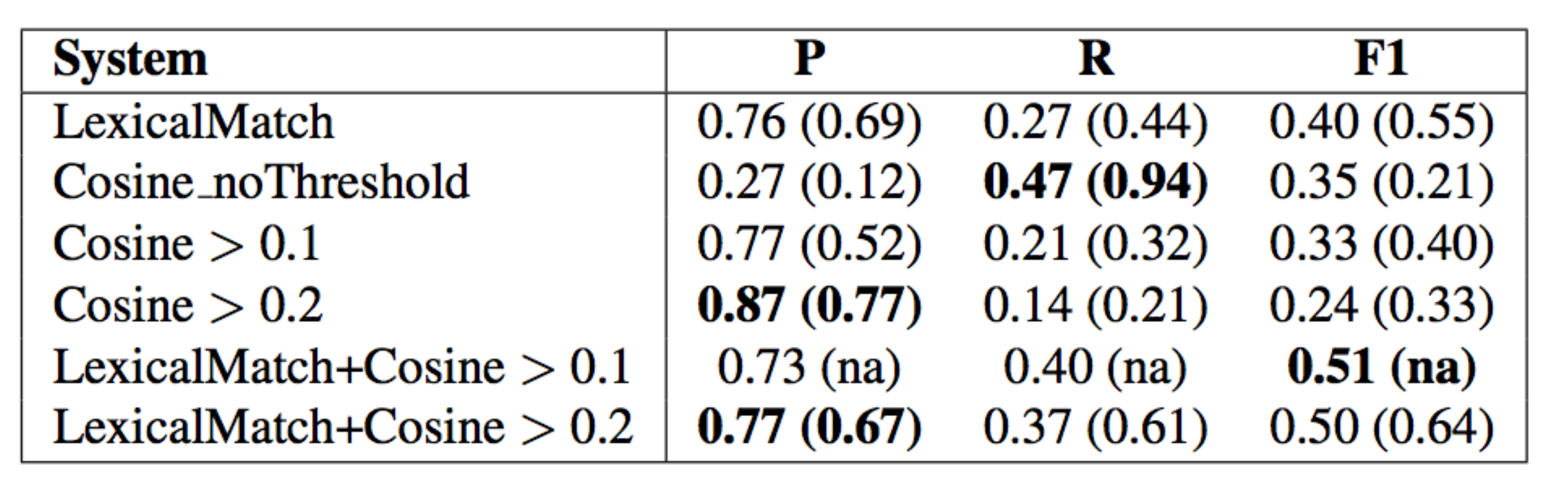
Evaluating the impact of Domain Information on WSA

Evaluation dataset: 166 aligned sense pairs of nouns (same data as in Caselli et al., 2014)

Lexical Match
WSA: token overlap between the MWN and SC sense descriptions.
Filtering: maximum overlap score + domain

Similarity Measure (Cosine)
WSA: cosine measure between vectors
Vectors obtained from Personalized Page Rank (UKB tool)
Knowledge Base: MWN and WN 3.0
Filtering: cut-off thresholds + domain

Merged (LexicalMatch+Cosine)
WSA: union of the the results obtained from Lexical Match and Similarity Measure.



System	P	R	F1
LexicalMatch	0.76 (0.69)	0.27 (0.44)	0.40 (0.55)
Cosine_noThreshold	0.27 (0.12)	**0.47 (0.94)**	0.35 (0.21)
Cosine > 0.1	0.77 (0.52)	0.21 (0.32)	0.33 (0.40)
Cosine > 0.2	**0.87 (0.77)**	0.14 (0.21)	0.24 (0.33)
LexicalMatch+Cosine > 0.1	0.73 (na)	0.40 (na)	**0.51 (na)**
LexicalMatch+Cosine > 0.2	**0.77 (0.67)**	0.37 (0.61)	0.50 (0.64)

EMNLP 2014: Conference on Empirical Methods in Natural Language Processing — October 25–29, 2014 — Doha, Qatar.