Natural Language Engineering **8** (4): 279–291.
DOI: 10.1017/S1351324902002966 Printed in the United Kingdom

Introduction to the special issue on evaluating word sense disambiguation systems

PHILIP EDMONDS
Sharp Laboratories of Europe, Oxford Science Park, Oxford OX4 4GB, UK
e-mail: phil@sharp.co.uk

ADAM KILGARRIFF
Information Technology Research Institute, University of Brighton, Lewes Road, Brighton BN2 4GJ, UK
e-mail: Adam.Kilgarriff@itri.brighton.ac.uk

- Definición:

  Word sense disambiguation (WSD) is the problem of deciding which sense a word has in any given context.

- Ejemplo:

$$\text{drug} \begin{cases} \rightarrow \text{medicamento} \\ \rightarrow \text{droga} \end{cases}$$

- Bibliografia:

  - Ide and Véronis (1998)
  - Manning and Schütze (1999)
  - Jurafsky and Martin (2000)

EXERCISES

- Recuperación de información → TREC
- Extracción de información → MUC
- Resumen automático → DUC
- Desambiguación de sentidos → SENSEVAL (ACL)
  - SENSEVAL I (1998, Inglés, francés e italiano)
  - SENSEVAL II (2000-2001, 3 tareas en 13 lenguas)
  - SENSEVAL III (2004 en Barcelona –ACL-04-)
  - http://www.senseval.org/

DARPA

A typology of evaluations

Sense inventory	*In vitro* evaluation	*In vivo* evaluation
Explicit application-independent	SENSEVAL	?
Explicit, defined by an application or domain	E.g. senses as translation equivalents	E.g. improvement in machine translation or information extraction as the task
Implicit, defined by application in a domain	E.g. senses as word or context clusters	E.g. improvement of information retrieval as the task

SENSEVAL Tasks

- **all-words** task, systems must tag almost all of the content words in a sample of running text.
- **lexical sample** task, we first carefully select a sample of words from the lexicon; systems must then tag several instances of the sample words in short extracts of text.
- **translation** task (Japanese only) is a lexical sample task in which word sense is defined according to translation distinction.
- http://www.sle.sharp.co.uk/senseval2/archive/call-for-participation.txt

Results of SENSEVAL-2

Table 3. *Results of* SENSEVAL-2, *tabulated from Edmonds and Cotton (2001)*

Language	Task[a]	Systems	Lemmas	Instances[b]	IAA[c]	Baseline[d]	Best score
Czech	AW	1	–[e]	277,986	–	–	0.94
Basque	LS	3	40	5,284	0.75	0.65	0.76
Dutch[f]	AW	1	1,168	16,686	–	0.75	0.84
English	AW	21	1,082	2,473	0.75	0.57	0.69
English	LS	26	73	12,939	0.86	0.48/0.16[g]	0.64/0.40
Estonian	AW	2	4,608	11,504	0.72	0.85	0.67
Italian	LS	2	83	3,900	0.21	–	0.39
Japanese	LS	7	100	10,000	0.86	0.72	0.78
Japanese	TL	9	40	1,200	0.81	0.37	0.79
Korean	LS	2	11	1,733	–	0.71	0.74
Spanish	LS	12	39	6,705	0.64	0.48	0.65
Swedish	LS	8	40	10,241	0.95	–	0.70

SEMCOR corpus

```
<contextfile concordance=brown>
<context filename=br-l11 paras=yes>
<s snum=3>
<wf pos=PRP>He</wf>
<wf pos=VB lemma=demonstrate wnsn=2 lexsn=2:31:00::>demonstrated</wf>
<wf pos=IN>by</wf>
<wf pos=VB lemma=play wnsn=7 lexsn=2:36:01::>playing</wf>
<wf pos=DT>an</wf>
<wf pos=JJ lemma=imaginary wnsn=1 lexsn=5:00:00:unreal:00>imaginary</wf>
<wf pos=NN lemma=piano wnsn=1 lexsn=1:06:00::>piano</wf>
<punc>,</punc>
<wf pos=VB lemma=do wnsn=2 lexsn=2:36:01::>doing</wf>
<wf pos=DT>a</wf>
<wf pos=JJ lemma=staccato wnsn=1 lexsn=3:00:00::>staccato</wf>
<wf pos=NN lemma=passage wnsn=6 lexsn=1:10:01::>passage</wf>
<wf pos=IN>with</wf>
<wf pos=DT>a</wf>
<wf pos=RB lemma=broadly wnsn=1 lexsn=4:02:00::>broadly</wf>
<wf pos=JJ lemma=exaggerated wnsn=1 lexsn=5:00:00:immoderate:00>exaggerated</wf
<wf pos=NN lemma=attack wnsn=8 lexsn=1:04:01::>attack</wf>
<punc>.</punc>
</s>
...
```

English lexical sample

Table 1. *Keyword-itemized performance on* SENSEVAL2 *English lexical sample task*

Model	Num Samples	Num Senses	ML	Entr	FENBayes	BayesRatio	Cosine	DL	TBL
begin.v	557	8	59.1%	0.2	79.4%	79.2%	80.3%	81.3%	**83.1%**
call.v	132	23	25.7%	0.5	**43.9%**	38.6%	35.6%	39.4%	40.2%
carry.v	132	27	23.5%	0.6	37.9%	**43.2%**	43.2%	39.4%	40.1%
collaborate.v	57	2	91.2%	0.1	86.1%	**94.7%**	87.9%	91.2%	**94.7%**
develop.v	133	15	30.1%	0.5	36.9%	38.4%	**41.3%**	40.6%	36.0%
art.n	196	19	38.2%	0.4	59.7%	65.9%	63.8%	61.7%	**67.3%**
authority.n	184	11	33.7%	0.3	**69.1%**	69.0%	64.1%	60.4%	66.4%
bar.n	304	22	41.8%	0.4	**71.4%**	71.0%	69.4%	63.1%	65.1%
bum.n	92	6	70.6%	0.3	69.5%	70.6%	62.0%	71.8%	**73.9%**
chair.n	138	8	82.6%	0.2	**91.3%**	91.3%	88.4%	89.9%	88.4%
channel.n	145	10	40.7%	0.4	60.0%	**62.1%**	61.4%	49.7%	48.3%
child.n	129	9	60.4%	0.2	68.2%	66.6%	64.3%	72.1%	**78.2%**
blind.a	108	9	62.9%	0.3	**74.2%**	71.5%	70.5%	72.3%	72.2%
colourless.a	68	3	77.9%	0.2	80.9%	**82.4%**	82.3%	77.8%	81.0%
cool.a	106	8	50.0%	0.4	**70.7%**	59.4%	52.9%	66.1%	56.6%
faithful.a	47	3	72.2%	0.2	65.6%	67.8%	59.6%	**74.2%**	70.0%

Senses for ‘bank’

Table 1. WordNet *senses and domains for the word ‘bank’*

Sense	Synset and Gloss	Domains	Semcor
#1	depository financial institution, bank, banking concern, banking company (a financial institution . . .)	ECONOMY	20
#2	bank (sloping land . . .)	GEOGRAPHY, GEOLOGY	14
#3	bank (a supply or stock held in reserve . . .)	ECONOMY	–
#4	bank, bank building (a building . . .)	ARCHITECTURE, ECONOMY	–
#5	bank (an arrangement of similar objects . . .)	FACTOTUM	1
#6	savings bank, coin bank, money box, bank (a container . . .)	ECONOMY	–
#7	bank (a long ridge or pile . . .)	GEOGRAPHY, GEOLOGY	2
#8	bank (the funds held by a gambling house . . .)	ECONOMY, PLAY	
#9	bank, cant, camber (a slope in the turn of a road . . .)	ARCHITECTURE	–
#10	bank (a flight maneuver . . .)	TRANSPORT	–

Lexical entry for noun headword "arte"

arte#NCMS#1#Actividad humana o producto de tal actividad que expresa simbólicamente un aspecto de la realidad: el arte de la música; el arte precolombino#SIN:?#00518008n/02980374n#
arte#NCMS#2#Sabiduría, destreza o habilidad de una persona en una actividad o conducta determinada: tiene mucho arte bailando; desplegó todo su arte para convencerle#SIN:?#03850627n#
arte#NCMS#3#Aparato que sirve para pescar#SIN:?#02005770n#

The papers

- Uso de la información de dominio en WSD
- Combinación de clasificadores
- Análisis del espacio de rasgos
- Ajuste de parámetros
- Método semisupervisado para WSD
- Evaluación de recursos léxicos

Natural Language Engineering **8** (4): 359–373.
DOI: 10.1017/S1351324902003029 Printed in the United Kingdom

The role of domain information in Word Sense Disambiguation

BERNARDO MAGNINI, CARLO STRAPPARAVA, GIOVANNI PEZZULO and ALFIO GLIOZZO
ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica, I-38050 Trento, Italy
e-mail: {magnini,strappa,pezzulo,gliozzo}@itc.it

(*Received 1 November 2001; revised 25 July 2002*)

Introduction

"*The purpose of this paper is to investigate the role of domain information in Word Sense Disambiguation.*

The hypothesis is that domain labels (such as Medicine, Architecture and Sport) provide a powerful way to establish semantic relations among word senses, which can be probably used during the disambiguation process."

chair_1: (F) a seat for one person,...
chair_2: (a) the position of professor
chair_3: (b) the officer who presides at the meeting .
chair_4: (c) an instrument of death ...

F *?*

The dinnertable and chairs are elegant yet comfortable,

and you can be assured of the finest tableware and crystal for

F

meals at home.

WordNet domains

- Es una extensión de WordNet donde cada synset tiene asociado una o mas etiquetas de dominio.
- Construcción semimanual basado en criterios sintactico-semánticos (*is-a*, *part-of*, etc.)
- Jerarquia de dominios (200/43)
- FACTOTUM (38620):
  - Synsets genéricos (*man_1*)
  - Bloquear mecanismo de propagación en palabras frecuentes (números, dias de la semana, colores, etc.).

Domains & words

- *Text Related Domain words*: palabras con al menos un sentido en el dominio del texto. Ej 'bank' en Economía
- *Text Unrelated Domain words*: palabras con ningún sentido perteneciente al dominio del texto. Ej. 'churh'
- *Text Unrelated Generics*: palabras cuyo sentido no es significativo. Ej. 'be'. Sense: FACTOTUM

Quantitative distribution

Table 3. *Word distribution in Semcor according to the prevalent domains of the texts*

Word class	Nouns	Verbs	Adjectives	Adverbs	All
TRD words	18732 (34.5%)	2416 (8.7%)	1982 (9.6%)	436 (3.7%)	21%
Polysemy	3.90	9.55	4.17	1.62	4.46
TUD words	13768 (25.3%)	2224 (8.1%)	815 (3.9%)	300 (2.5%)	15%
Polysemy	4.02	7.88	4.32	1.62	4.49
TUG words	21902 (40.2%)	22933 (83.2%)	17987 (86.5%)	11131 (93.8%)	64%
Polysemy	5.03	10.89	4.55	2.78	6.39

One domain per discourse

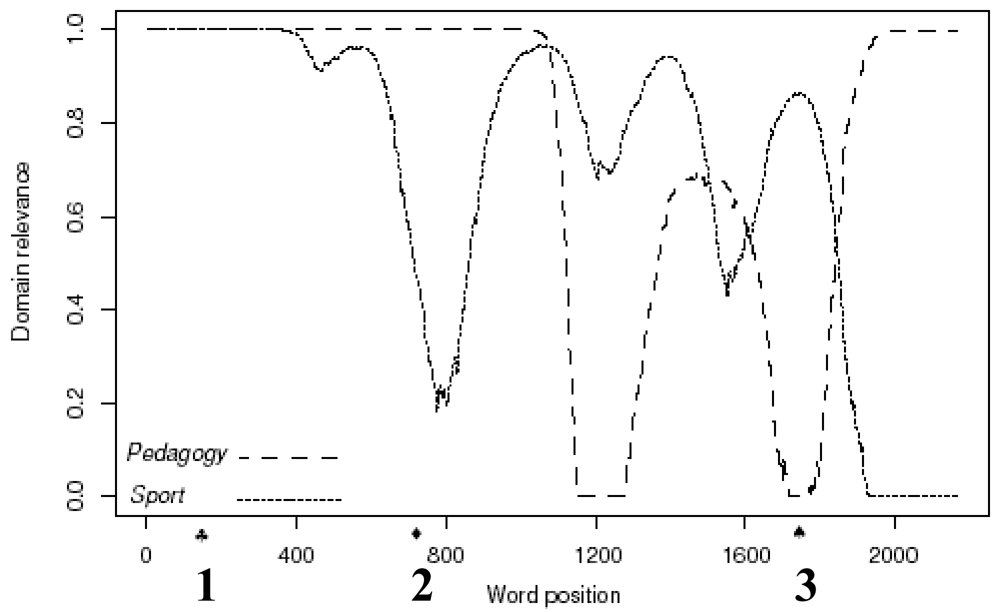
- *One Sense per Discourse hypothesis*: en textos bien escritos existe la tendencia a que los distintos usos de una palabra correspondan a un mismo sentido
  Gale, Church & Yarowsky (1992) vs Krovetz (1998)

- *One Domain per Discourse*: distintos usos de una misma palabra en fragmentos coherentes de texto corresponden a un mismo domino

Table 4. *One Sense per Discourse vs. One Domain per Discourse*

Pos	Cases[a]	Exceptions to OSD[b]	Exceptions to ODD[c]
All	23877	7469 (31%)	2466 (10%)
Nouns	10291	2403 (23%)	1142 (11%)
Verbs	6658	3154 (47%)	916 (13%)
Adjectives	4495	1100 (24%)	391 (9%)
Adverbs	2336	790 (34%)	12 (1%)[d]

Domain variation in a text



1. The Russians are all trained as dancers before they start to study gymnastics . .
2. If we wait until children are in junior-high or high-school, we will never manage it. . . .
3. The backbend is of extreme importance to any form of free gymnastics, and, as with all acrobatics, the sooner begun the better the results. . . .

Domains and WSD

- Metodologia básica: comparación entre los dominios de la palabra a desambiguar y los dominios de las palabras del contexto.
- Estructura de dades: *domain vector*
  - *text vector* relevancia de un fragmento en relación a cada dominio
  - *domain vector* relevancia de cada sentido de cada palabra en relación a cada dominio

Domain relevance

- Número positivo [0,1]
- Cálculo:
  - Define un ventana *v* (≥ 25)
  - Calcula la frecuencia de los dominios
  - Compara el resultado con un corpus balanceado (LOB) suponiendo una distribución normal
- Ej. "*Today I draw money from my bank*"

1/33 + 5/10 + 3/3 = 1,53 > 0,4

LOB (ECONOMY) → 0,2 · σ = 0.1

Text vector & Sense vector

- Text vector: es un vector de dominio calculado a partir de un fragmento de texto
  - Dados T, p y $\{D_1, D_2,...D_n\}$ → $\vec{T}p$

- Sense vector: es un vector de dominio calculado a partir de los sentidos de una palabra.

  Ej. bank_1 → (economy·20, sport·0, ...)

Disambiguation procedure

- Comparación entre el *Text vector* y el *Sense vector* (todos los sentidos)

- Ejemplo

Table 5. *Sense vectors ($\vec{s}_1$ and $\vec{s}_2$) and text vector ($\vec{T}_8$) for the text* T *'Today I have drawn money from my bank', for a subset of domains*

	Sport	Medicine	Economy	Geography
$\vec{s}_1$ (*Bank#1*)	0.02	0.08	1.73	0.04
$\vec{s}_2$ (*Bank#2*)	0.005	0.03	0.04	0.69
$\vec{T}_8$	0.2	0.005	1	0.03

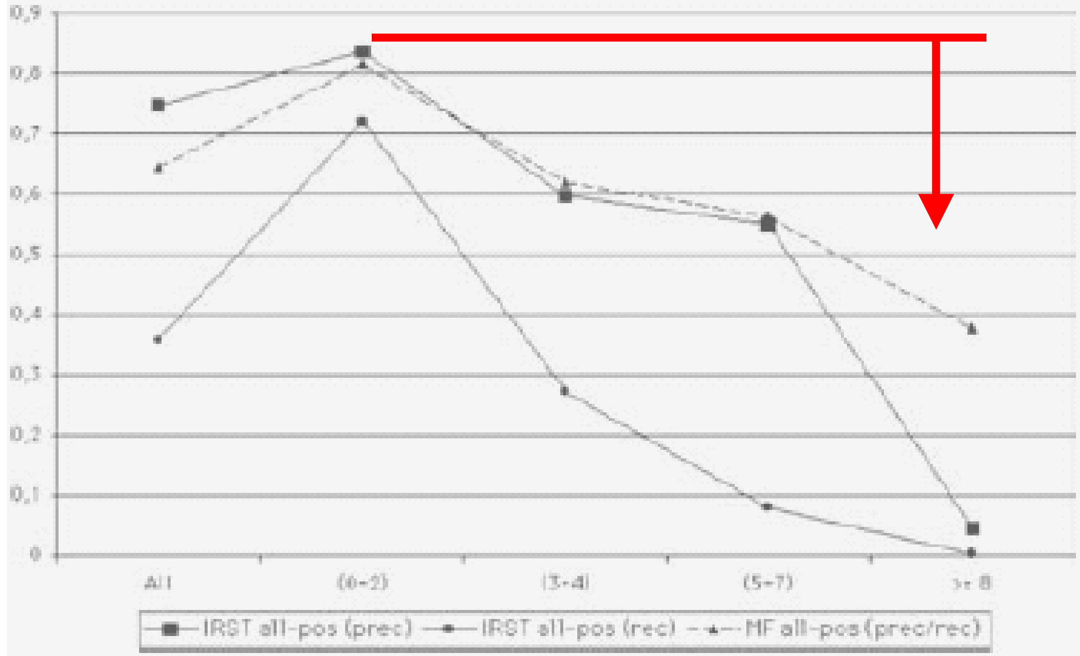
Ts · s1 = 1,73

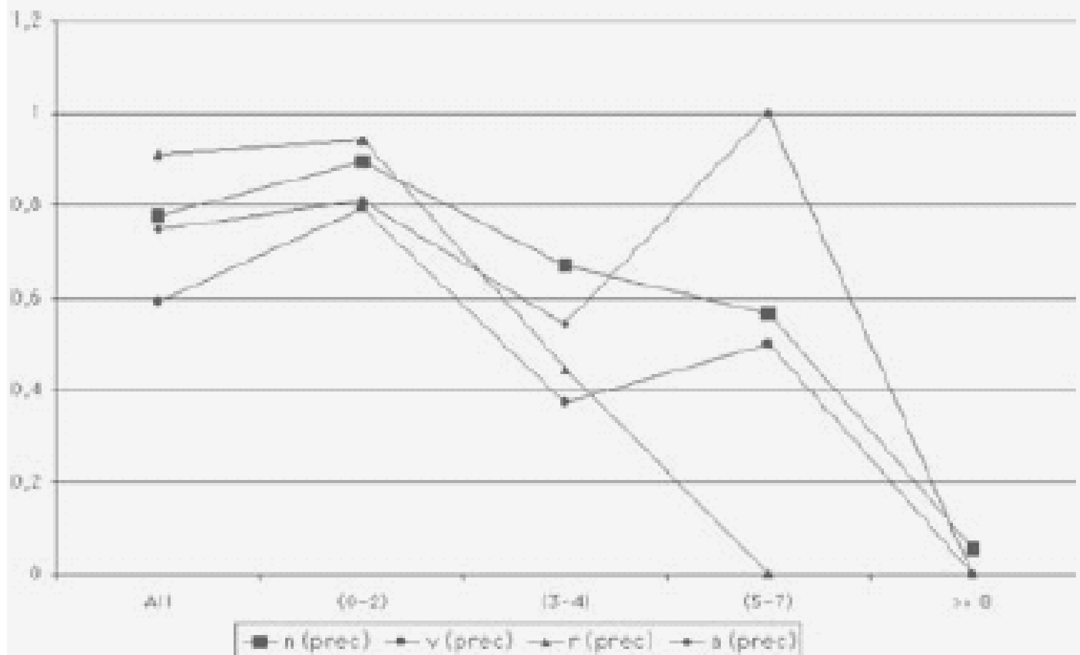
Ts · s2 = 0.06

Results

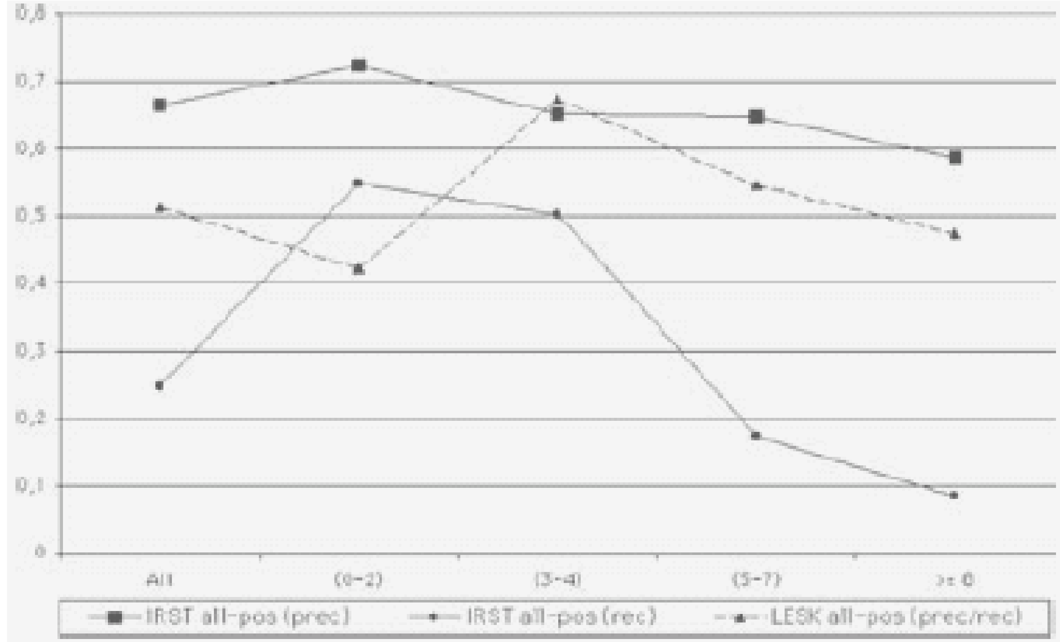
- El sistema participó en dos tareas:
  - *English_all_words* (desambiguar todas las palabras de un texto)
  - *English_lexical_sample* (una palabra y su contexto)
- ds

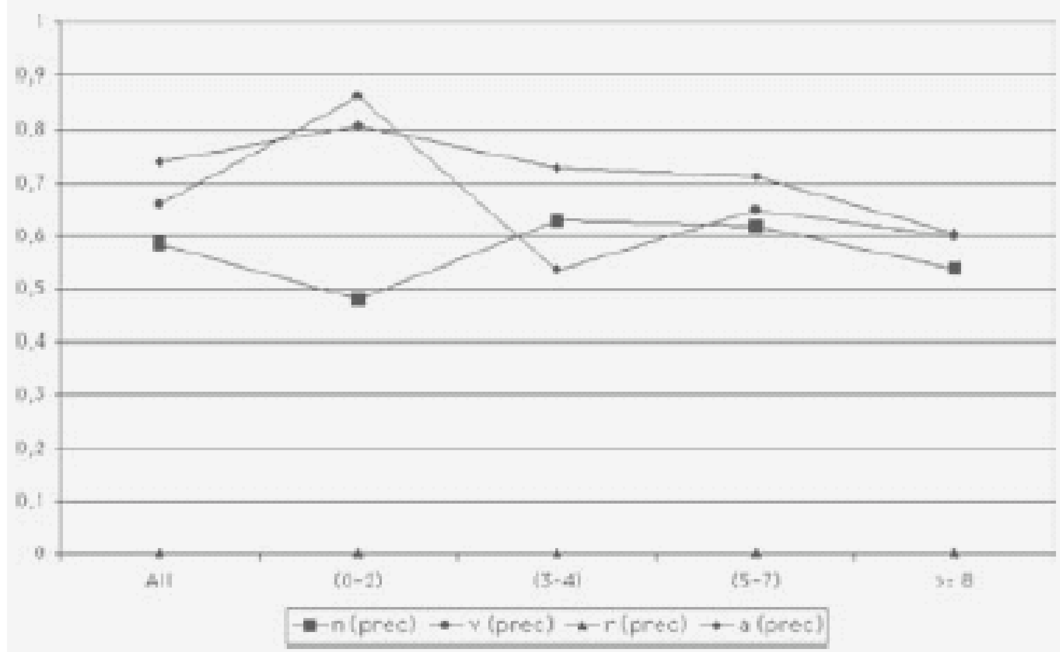
all_words Task





lexical_sample_words Task





Conclusions

- El algoritmo de desambiguación propuesto aprovecha la información de dominio y
- Para un subconjunto importante de palabras obtiene un alto grado de precisión .

Natural Language Engineering **8** (4): 327–341.
DOI: 10.1017/S1351324902002978 Printed in the United Kingdom

Combining classifiers for word sense disambiguation

RADU FLORIAN, SILVIU CUCERZAN,
CHARLES SCHAFER and DAVID YAROWSKY
Department of Computer Science and Center for Language and Speech Processing
Johns Hopkins University, MD 21218, USA
e-mail: {rflorian,silviu,cschafer,yarowsky}@cs.jhu.edu

(*Received 19 November 2001; revised 20 June 2002*)

Introduction

- La combinación de clasificadores es una manera de mejorar el rendimiento de ciertas aplicaciones
- Cada método tiene sus puntos fuertes y funciona bien sobre diferentes tipos de datos de prueba. Existen:
  - Características inherentes a cada método
  - Diferencias en los métodos de selección de los rasgos
  - Uso de diferentes fuentes de conocimiento en la fase de entrenamiento.
- Métodos aprendizaje (supervisado): *Naïve Bayes* (variantes), *Cosine* y *Decision lists*
- Tarea: *lexical-sample task*
- Idiomas: inglés, español, vasco y sueco

The feature space

- Aspecto crítico en el diseño de un clasificador
- Rasgos morfológicos: lema + categoría
- Rasgos sintácticos :
  - Verbos: núcleo nominal del objeto, preposición y PP
  - Nombres: función sintáctica (sujeto, objeto, ...)
  - Adjetivos: núcleo nominal que modifica

Example sentence

Table 1. *Example sentence and sample of extracted features*

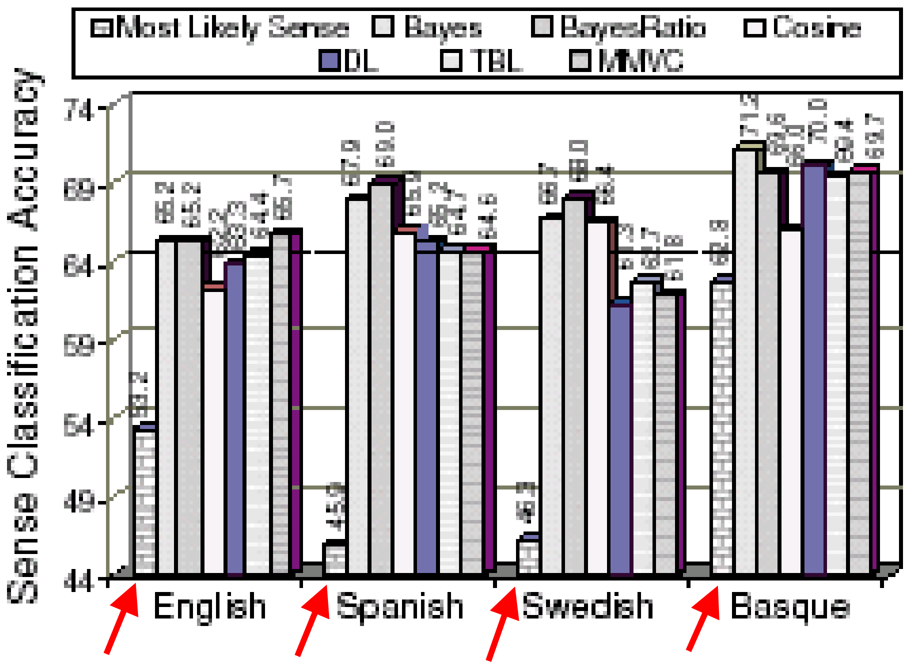
Many mothers do not even try to toilet **train** their children until the age of 2 years or later ..

Feature type	Word	POS	Lemma	Feature type	Word	POS	Lemma
Context	...	...	...	*Syntactic (predicate-argument) features*			
Context	try	VB	try/N	Object	children	NNS	child/N
Context	to	TO	to/T	Prep	until	IN	until/I
Context	toilet	NN	toilet/N	ObjPrep	age	NN	age/N
Context	train	VBP	train/V	*Ngram collocational features*			
Context	their	DT	their/D	−1 bigram	toilet	NN	toilet/N
Context	children	NN	child/N	+1 bigram	their	DT	their/D
Context	...	...	...	−1/+1 trigram	to * their	TO-DT	to/T * their/D
Context	...	...	...	+1/+2 trigram	their children	DT-NN	their/D child/N

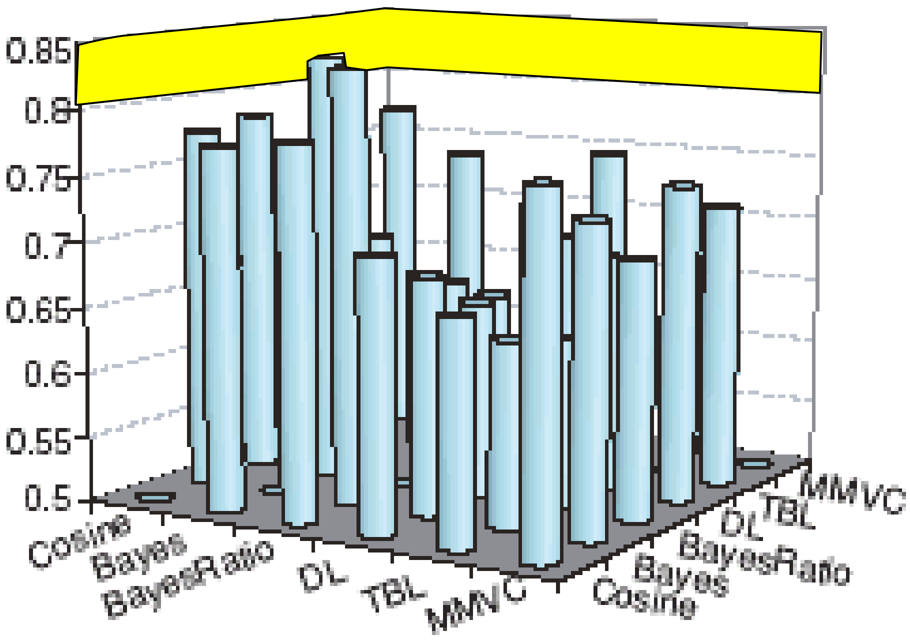
WSD classifier combination

- Classifer combination has been theoretically and practically shown to be beneficial in terms of improving system accuracy. Perrone (1993) shows that, under the restrictive assumption that the n input classifers are uncorrelated and have unbiased binary output, the expected error is reduced by a factor of n when combining their classications through averaging.

Individual classifier properties

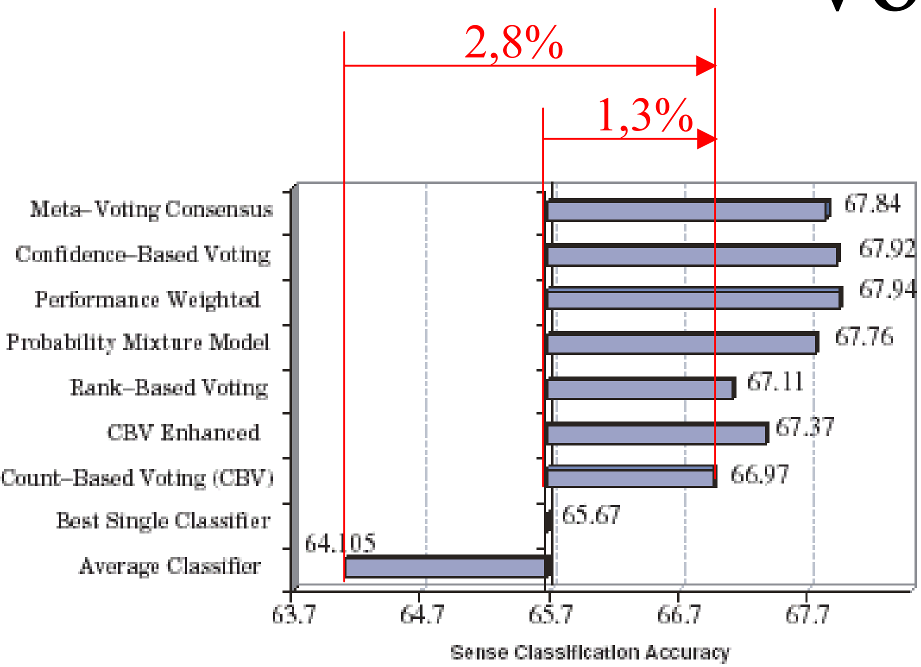


(a) Individual classifer performance,

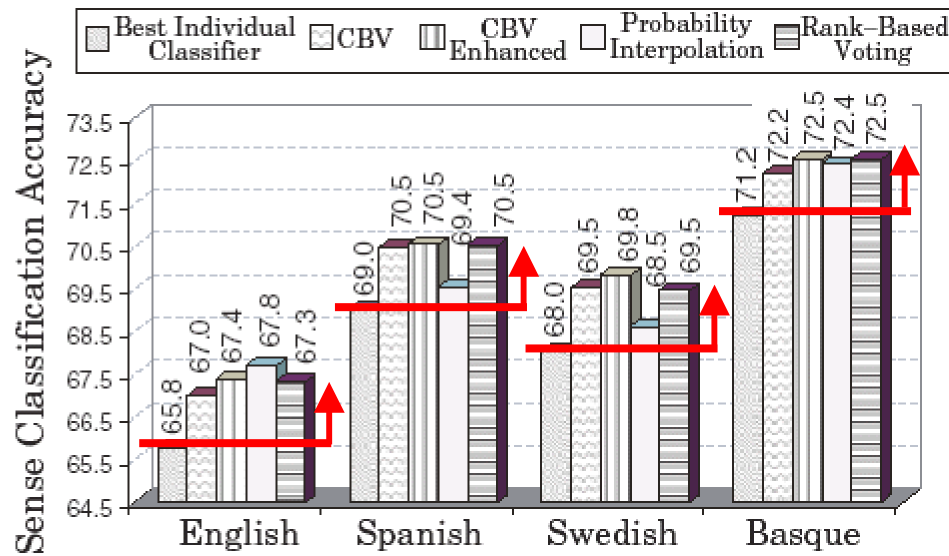


(b) Classifier inter-agreement.

Count-based and probability-based voting



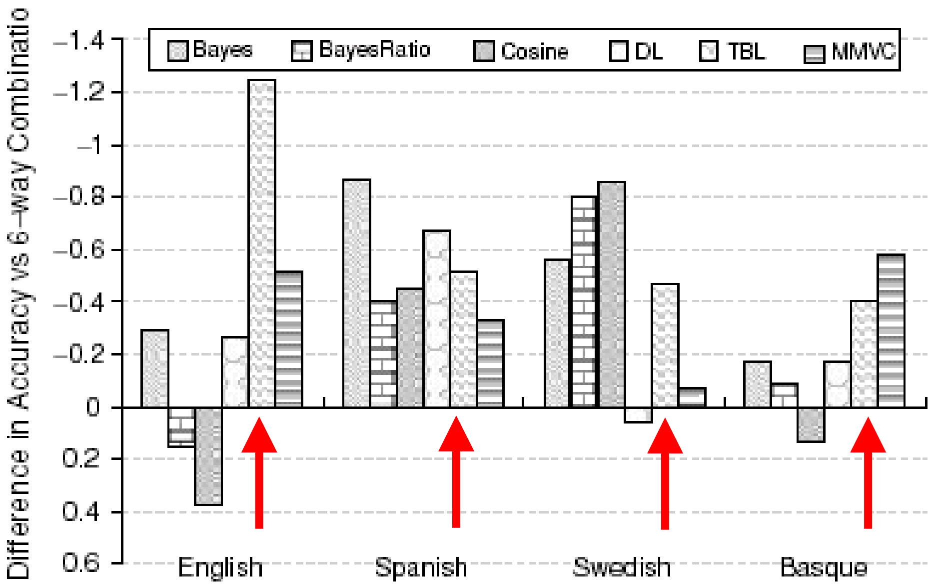
(a) English lexical choice WSD performance,



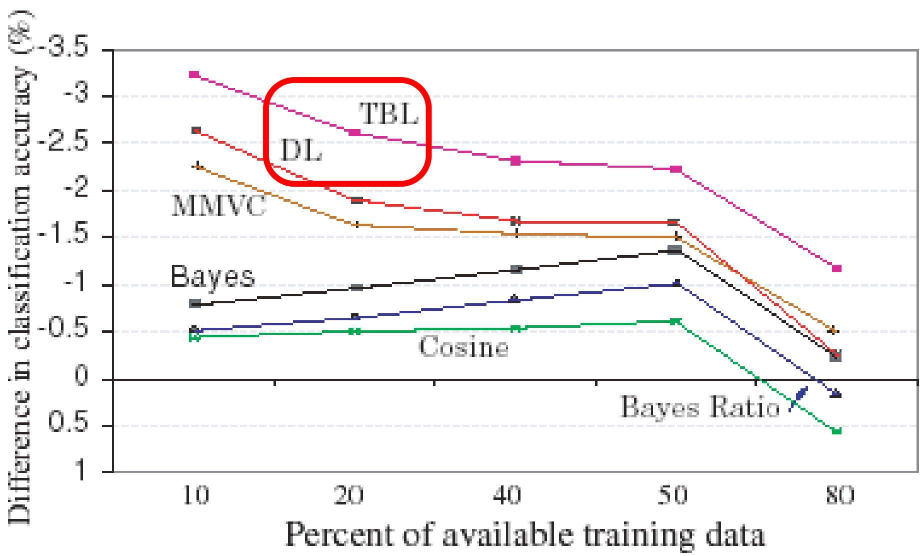
(b) WSD performance across four languages.

CBV enhanced= CBV + probabilidades
Probability mixture= voto ponderado
Rank Based Voting= voto ponderado
Metavoting= CBV sobre los *k* mejores métodos

Individual classifers' contribution to combination



(a) Performance drop when eliminating one classifer



(b) Performance when eliminating one classifer, by training data size.

Natural Language Engineering **8** (4): 293–310. © 2002 Cambridge University Press 293
DOI: 10.1017/S135132490200298X Printed in the United Kingdom

Evaluating sense disambiguation across diverse parameter spaces

DAVID YAROWSKY and RADU FLORIAN
Department of Computer Science and Center for Language and Speech Processing, Johns Hopkins University, MD 21218, USA
e-mail: {yarowsky,rflorian}@cs.jhu.edu

(*Received 19 November 2001; revised 20 June 2002*)

Introduction

- Es un estudio comparativo y detallado de la influencia que tienen algunos parámetros utilizados en diferentes algoritmos de WSD
- Algoritmos: variantes de *Naïve Bayes*, *cosine model*, *TBL* y *decision lists*
- Parámetros analizados:
  - Idioma
  - POS
  - Granularidad
  - Ancho de contexto
  - Núm. de ejemplos de entranamiento
  - Entropia en la distribución de los sentidos
  - Efecto del ruido
  - Divergencia entre datos de entrenamiento/prueba

Algorithm classification

- Aggregative:

  integrate all available evidence in favor of a sense and then select the sense with the maximum cumulative support (*cosine*, *FENBayes* and *BR*)

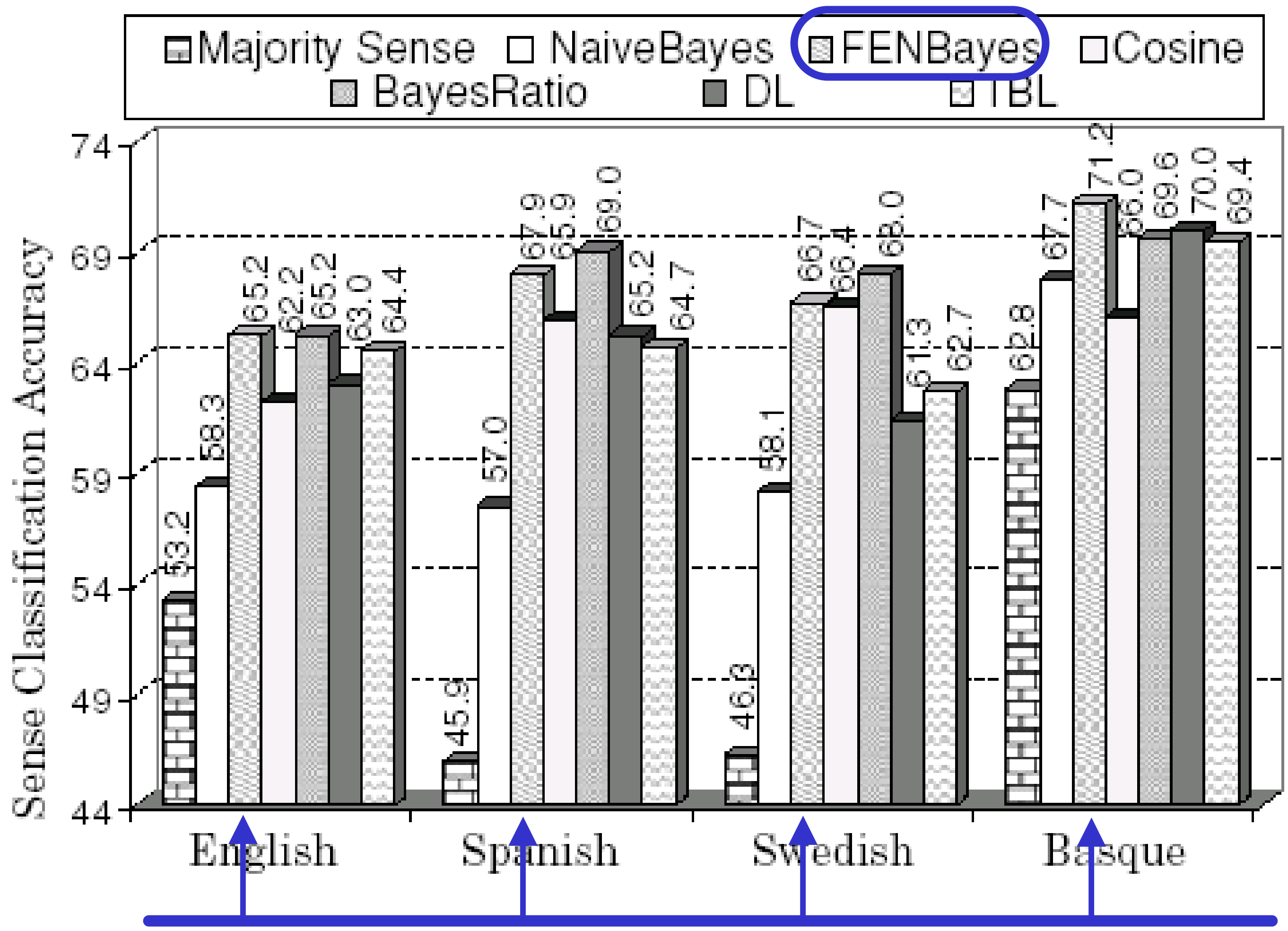
- Discriminative:

  rely on one or a few features in any given context that most efficiently partition or discriminate the candidate sense space (*DL* and *TBL*).

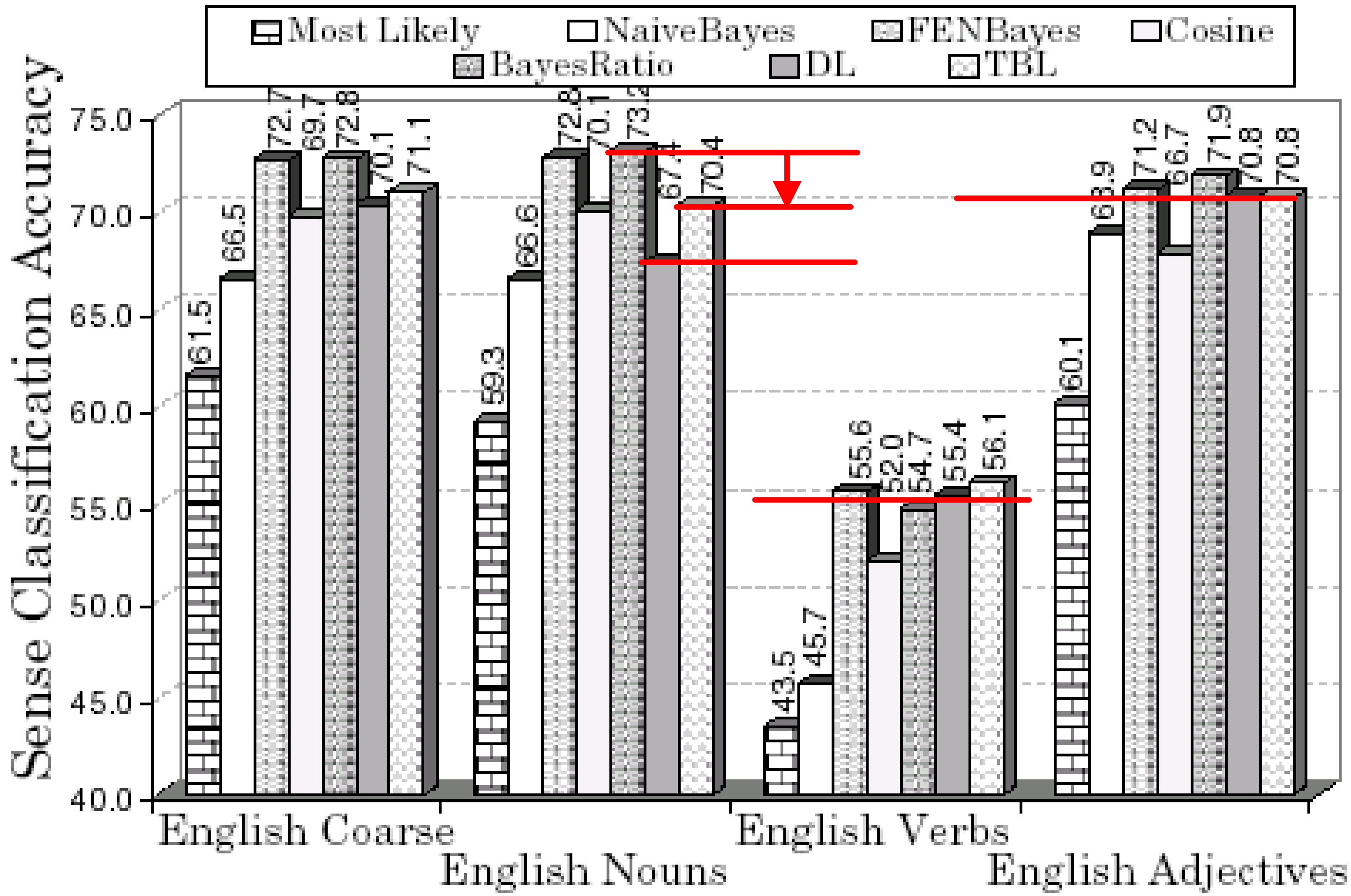
Feature space

- Forma/lema/etiqueta para conjuntos de palabras o colocaciones tipo n-gram.
- Relaciones sintácticas relevantes (sujeto, objeto, modificador, etc.)
- Agrupaciones:
  - *BagOfWordsContext*
  - *LocalContext*
  - *SyntacticFeatures*

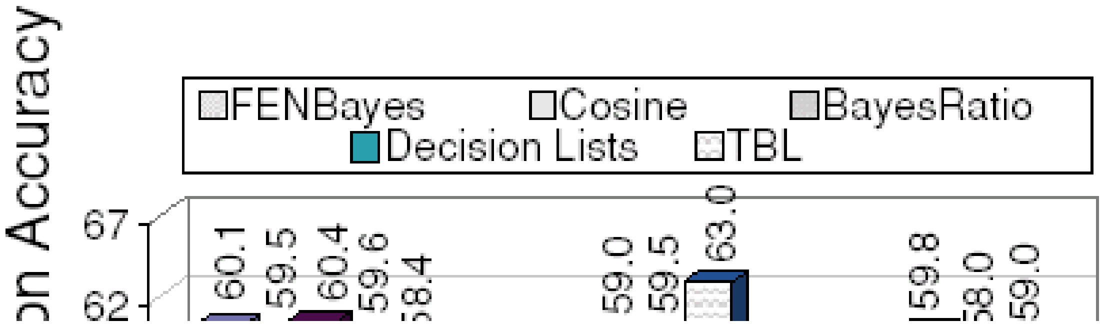
Performance based on language



Performance based on POS



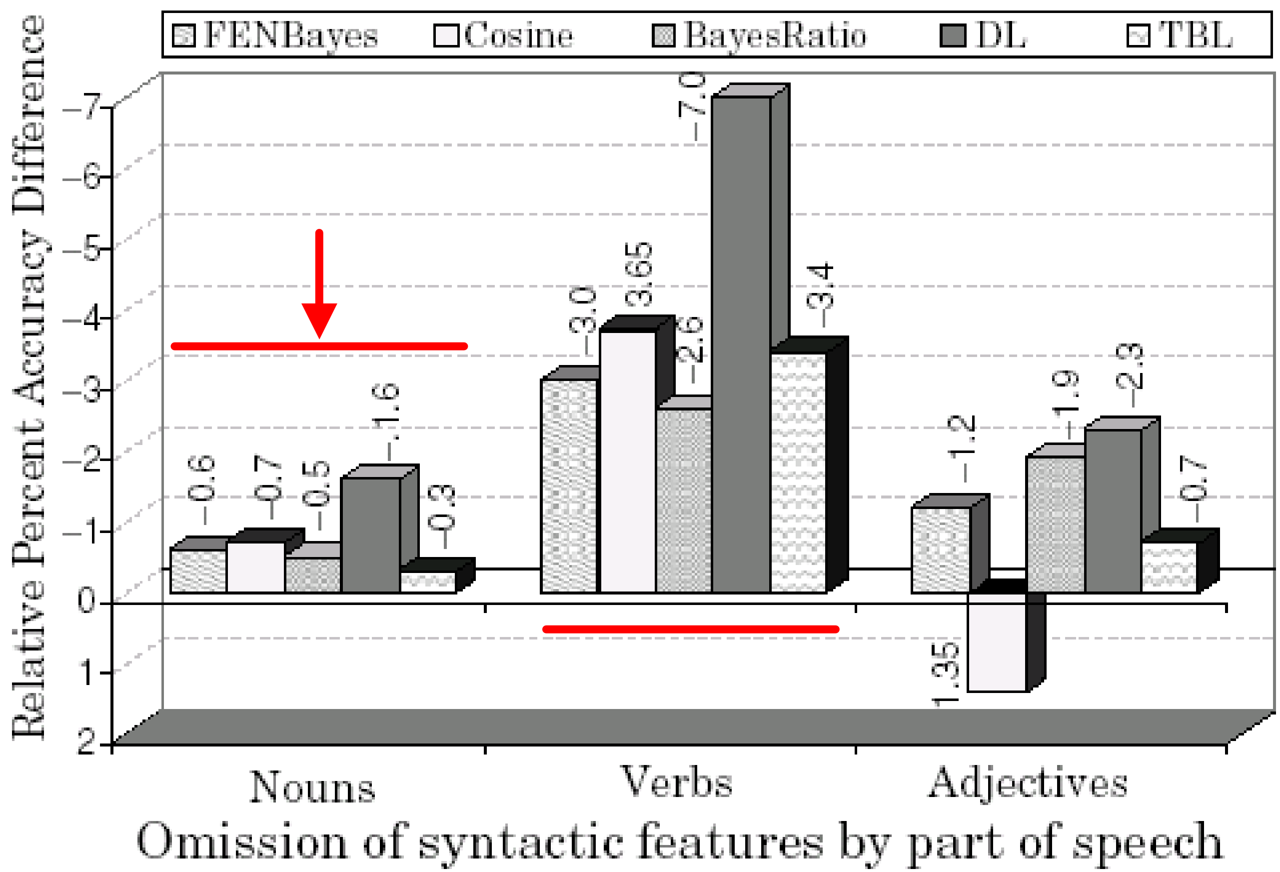
Performance sensitivity to feature type



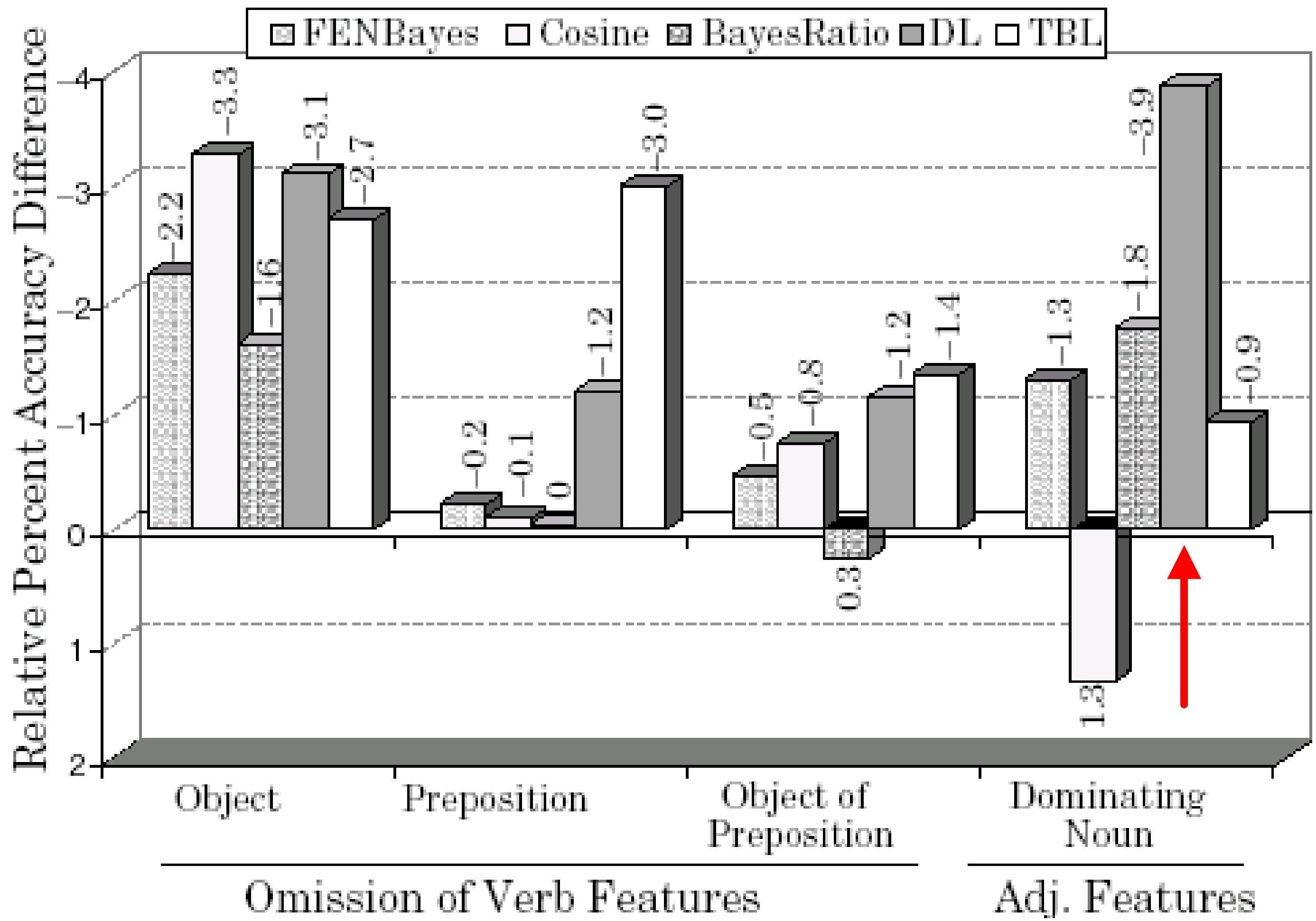
Accuracy drop relative to full system

Features used	Aggregative			Discrimin.	
	FENB	CSN	BR	TBL	DL
Omit Bag-of-words Ftrs	−14.7	−8.1	−5.3	−0.5	−2.0
Omit Local Collocations	−3.5	−0.8	−2.2	−2.9	−4.5
Omit Syntactic Features	−1.1	−0.8	−1.3	−1.0	−2.3
Bag-of-words Ftrs ***Only***	−6.4	−4.8	−4.8	−6.0	−3.2
Local Collocations ***Only***	−18.4	−11.5	−6.1	−1.5	−3.3
Syntactic Features ***Only***	−28.1	−14.9	−5.4	−5.4	−4.8

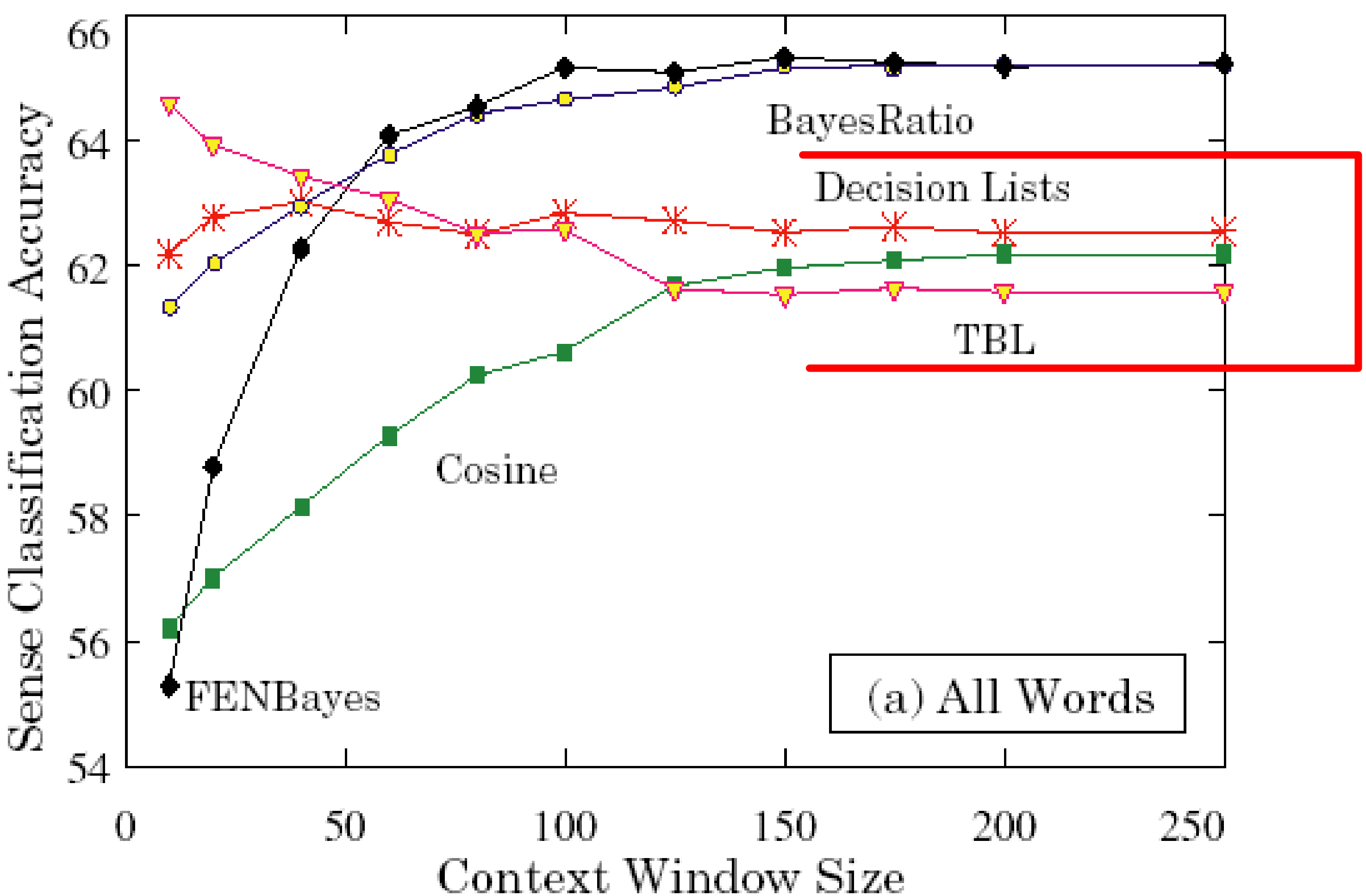
Contribution of syntactic features (I)



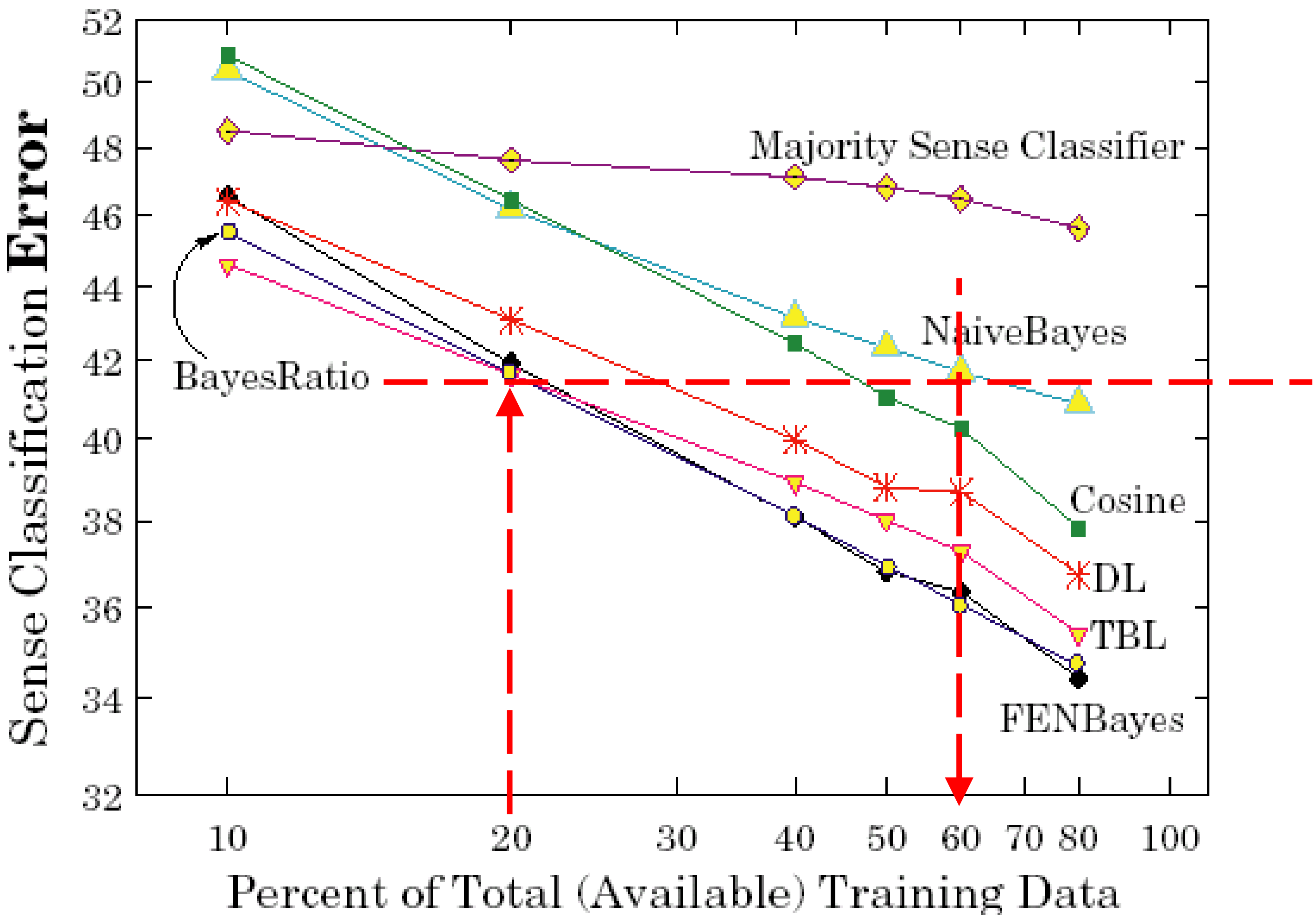
Contribution of syntactic features (II)



Performance sensitivity to context window size



Performance sensitivity to size of training data



Conclusions

- Ningún algoritmo destaca sobre el resto.

Model	Samples	Senses	ML	Entr	FENBayes	BayesRatio	Cosine	DL	TBL
wander.v	100	4	**83.0%**	0.1	78.0%	79.0%	63.0%	81.0%	82.0%
wash.v	25	13	8.0%	0.8	52.0%	52.0%	56.0%	**68.0%**	40.0%
work.v	119	21	27.6%	0.5	44.6%	**46.3%**	40.4%	40.5%	39.6%

- Los algoritmos discriminativos y agregativos tienen comportamientos complementarios → uso de algoritmos de combinación de clasificadores.
- La calidad del espacio de rasgos puede tener un impacto superior al de la elección del algoritmo de desambiguación

Natural Language Engineering **8** (4): 279–291

Introduction to the special issue on evaluating word sense disambiguation systems

PHILIP EDMONDS
Sharp Laboratories of Europe, Oxford Science Park, Oxford OX4 4GB, UK
e-mail: phil@sharp.co.uk
ADAM KILGARRIFF
Information Technology Research Institute, University of Brighton,
Lewes Road, Brighton BN2 4GJ, UK
e-mail: Adam.Kilgarriff@itri.brighton.ac.uk