Συνδυασμός τεχνικών Μηχανικής Μάθησης με στόχο τη μείωση των λαθών στο Μορφολογικό Σχολιασμό κειμένων

ΚΟΥΤΣΟΜΠΟΓΕΡΑ ΜΑΡΙΑ
ΚΩΝΣΤΑΝΤΙΝΙΔΗΣ ΑΛΕΞΗΣ
ΠΑΠΑΓΕΩΡΓΙΟΥ ΧΑΡΗΣ

ΙΝΣΤΙΤΟΥΤΟ ΕΠΕΞΕΡΓΑΣΙΑΣ ΤΟΥ ΛΟΓΟΥ (ΙΕΛ)

Μορφολογικός σχολιασμός
(Part-Of-Speech Tagging)

- Μονοσήμαντος χαρακτηρισμός των λεκτικών μονάδων ενός Σώματος Κειμένων βάσει ενός προκαθορισμένου συνόλου χαρακτηριστικών
- Ολοκλήρωση ΜΣ σε συστήματα ΕΦΓ
- Χρήση σχολιασμένου κειμένου (tagged text) σε άλλες εφαρμογές
- Επίπεδο **ακρίβειας**: 95-98%
- Προβληματικές περιπτώσεις
  - Άγνωστες λέξεις
  - Αμφίσημες λέξεις
  - Κλιτές γλώσσες

Δεδομένα

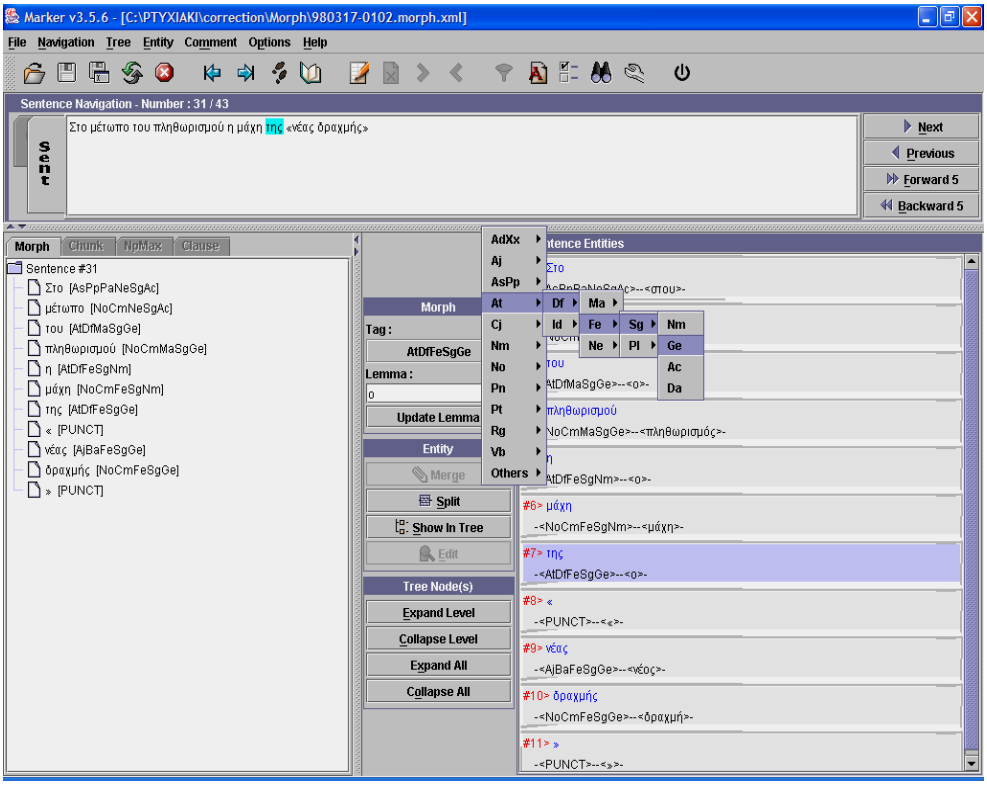
- Σώμα κειμένων
  - Σώμα εκπαίδευσης (700K)
  - Σώμα αξιολόγησης (23K)
- Λεξικό
- Σχήμα σχολιασμού
  - Σύνολο χαρακτηριστικών (47)
    - Βασική κατηγορία
      - NoCm, NoPr, CjCo, CjSb, AjBa, AjCo, AjSu…
  - Οδηγίες σχολιασμού
  - Σχήμα κωδικοποίησης
- Εργαλεία σχολιασμού

Δείγμα σχολιασμένου κειμένου

Πρόταση: *Ολόκληρη η εγκύκλιος και οι πλήρεις πίνακες για κάθε νομό*

<sent id="**s_7**" start="**125**" end="**182**">
<mw id="**mw_7_1**" lex="**Ολόκληρη**" tag="**AjBaFeSgNm**" lemma="**ολόκληρος**" start="**0**" end="**8**" />
<mw id="**mw_7_2**" lex="**η**" tag="**AtDfFeSgNm**" lemma="**ο**" start="**9**" end="**10**" />
<mw id="**mw_7_3**" lex="**εγκύκλιος**" tag="**NoCmFeSgNm**" lemma="**εγκύκλιος**" start="**11**" end="**20**" />
<mw id="**mw_7_4**" lex="**και**" tag="**CjCo**" lemma="**και**" start="**21**" end="**24**" />
<mw id="**mw_7_5**" lex="**οι**" tag="**AtDfMaPlNm**" lemma="**ο**" start="**25**" end="**27**" />
<mw id="**mw_7_6**" lex="**πλήρεις**" tag="**AjBaMaPlNm**" lemma="**πλήρης**" start="**28**" end="**35**" />
<mw id="**mw_7_7**" lex="**πίνακες**" tag="**NoCmMaPlNm**" lemma="**πίνακας**" start="**36**" end="**43**" />
<mw id="**mw_7_8**" lex="**για**" tag="**AsPpSp**" lemma="**για**" start="**44**" end="**47**" />
<mw id="**mw_7_9**" lex="**κάθε**" tag="**PnIdMa03SgAcXx**" lemma="**κάθε**" start="**48**" end="**52**" />
<mw id="**mw_7_10**" lex="**νομό**" tag="**NoCmMaSgAc**" lemma="**νομός**" start="**53**" end="**57**" />
</sent>

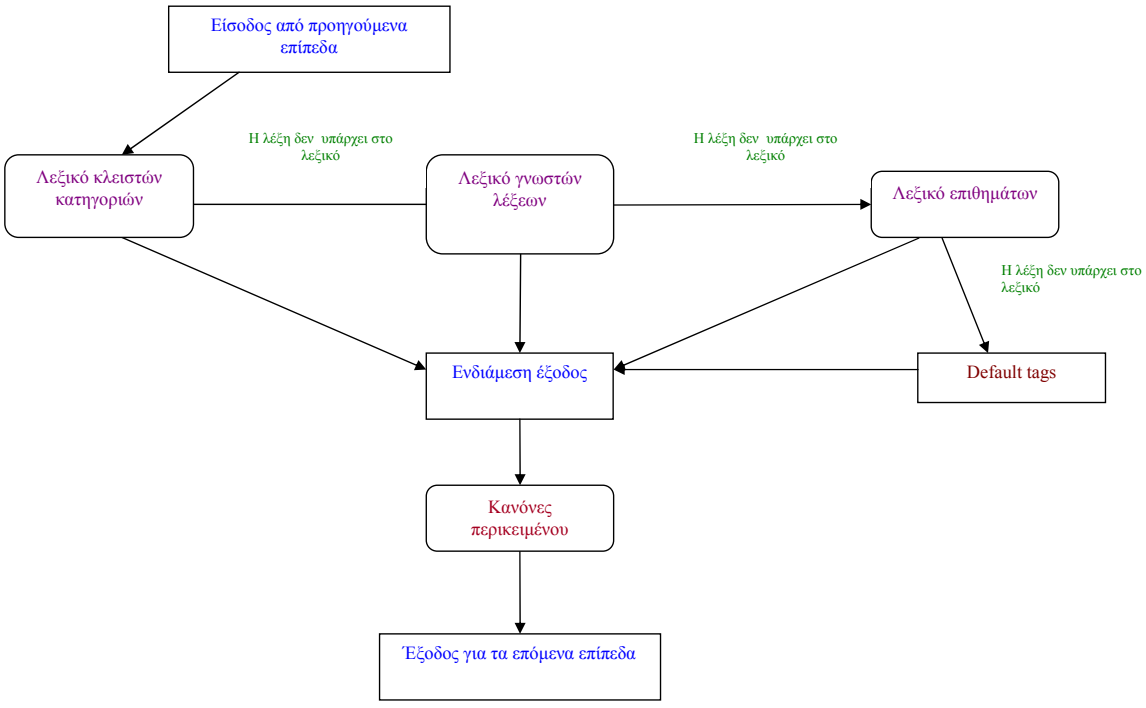
Παρουσίαση/ τροποποίηση σχολιασμένου κειμένου



Base Taggers

- Transformations
  - Feature-Based Tagging (FBT)
- Statistical approaches
  - Maximum Entropy (ME)
  - n-gram models (TnT)
- Memory-Based Learning
  - TiMBL

FBT


Είσοδος από προηγούμενα επίπεδα
Η λέξη δεν υπάρχει στο λεξικό
Η λέξη δεν υπάρχει στο λεξικό
Λεξικό κλειστών κατηγοριών
Λεξικό γνωστών λέξεων
Λεξικό επιθημάτων
Η λέξη δεν υπάρχει στο λεξικό
Ενδιάμεση έξοδος
Default tags
Κανόνες περικειμένου
Έξοδος για τα επόμενα επίπεδα

Maximum Entropy

- X: χαρακτηριστικά που αναφέρονται στο περικείμενο
- Y: χαρακτηριστικά που αναφέρονται στην υπό εξέταση λέξη

$$p^* = \arg\max_{p \in P} H(p)$$

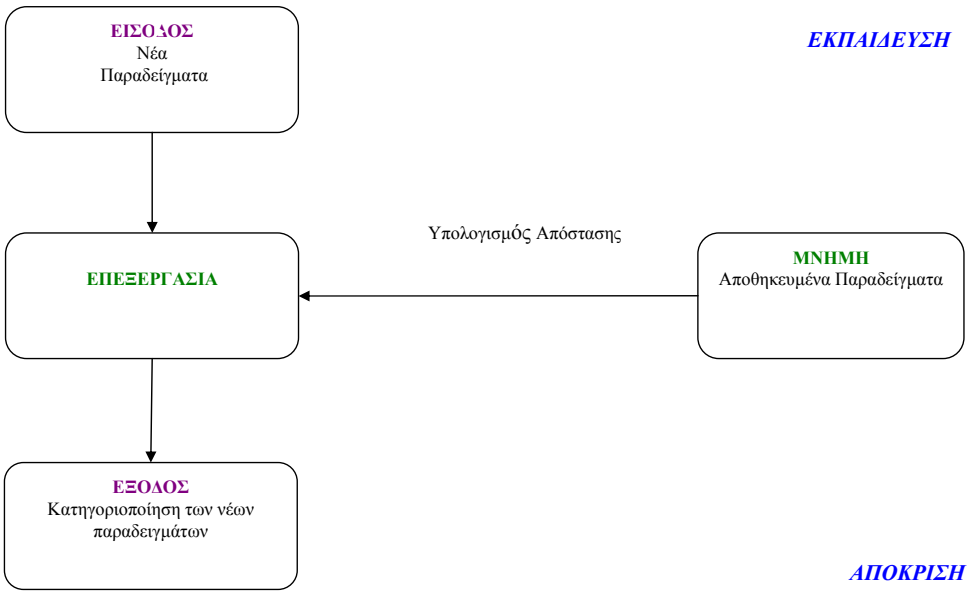
- Εκτίμηση των παραμέτρων:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{n} \lambda_i f_i(x, y)\right)$$

Maximum Entropy

- Αξιοποίηση της πληροφορίας του περικειμένου (1-3 λέξεις)
  - λέξη
  - λεξικό κλειστών κατηγοριών
  - λεξικό καταλήξεων
  - μορφολογικά χαρακτηριστικά (tags)
  - παρουσία ψηφίων, κεφαλαίων/λατινικών χαρακτήρων
  - σημεία στίξης
  - αριθμός τονούμενων συλλαβών
  - μήκος λέξης

MBT


ΕΙΣΟΔΟΣ
Νέα
Παραδείγματα
ΕΚΠΑΙΔΕΥΣΗ
Υπολογισμός Απόστασης
ΕΠΕΞΕΡΓΑΣΙΑ
ΜΝΗΜΗ
Αποθηκευμένα Παραδείγματα
ΕΞΟΔΟΣ
Κατηγοριοποίηση των νέων
παραδειγμάτων
ΑΠΟΚΡΙΣΗ

Memory-based tagging

- Χαρακτηριστικά γνωστών λέξεων
  - tags (0,±1)
- Χαρακτηριστικά αγνώστων λέξεων
  - 4 τελευταίοι χαρακτήρες w0
  - tags (±1)
- Αλγόριθμοι: IB1, IB1-IG
- Μετρικές: Overlap, Information Gain

HMM

- Καθορισμός ακολουθίας καταστάσεων (tags) που είναι πιο πιθανή για την παραγωγή συμβόλων εξόδου (words)
- $$\underset{T}{\text{argmax}}\, P(T|W) = \underset{T}{\text{argmax}}\, \frac{P(T)P(W|T)}{P(W)} = \underset{T}{\text{argmax}}\, P(T)P(W|T)$$
- TnT (Brants 2000)
- Trigrams: HMM μοντέλα β΄ τάξης
- Λεξική εγγραφή: πιθανά χαρακτηριστικά-συχνότητα εμφάνισης

Συνδυαστικές Τεχνικές Μορφολογικής Ανάλυσης

- Αξιοποίηση της συμπληρωματικότητας των Base Taggers
- Εκπαίδευση: 8-fold cross-validation
- Τύποι:
  - Majority Voting
  - Weighted Voting
  - Stacked Classifiers

Stacked Classifiers

- Εκπαίδευση Maximum Entropy Classifiers
  - Χαρακτηριστικά που προτείνονται από τους 4 base taggers
  - Περικειμενική πληροφορία: ±2 tags/words
  - Η λεξική πληροφορία δεν αποφέρει καλύτερα αποτελέσματα

Stacked Classifiers

lex=υπ' gold=AsPpSp				
model	**Tags (given)**	**Tags (all)**	**Tags**	**Tags (given)**
context	**-1, 0, +1**	**-1, 0, +1**	**0**	**-1, 0**
feature n_o	**9**	**12**	**4**	**5**
	t-1=NoCm t0_1=AdXxBa t0_2=NmOd t0_3=AsPpSp t0_4=AsPpSp t1_1=NoCm t1_2=NoCm t1_3=NoCm t1_4=NoCm	t-1_1=NoCm t-1_2=NoCm t-1_3=NoCm t-1_4=NoCm t0_1=AdXxBa t0_2=NmOd t0_3=AsPpSp t0_4=AsPpSp t1_1=NoCm t1_2=NoCm t1_3=NoCm t1_4=NoCm	t0_1=AdXxBa t0_2=NmOd t0_3=AsPpSp t0_4=AsPpSp	t-1=NoCm t0_1=AdXxBa t0_2=NmOd t0_3=AsPpSp t0_4=AsPpSp
accuracy	**98,244%**	**98,167%**	**98,180%**	**98,145%**

Αποτελέσματα

Tagger	FBT	ME	MBT	TnT	Majority	Weighted	Combiner
Accuracy (%)	97,91	**98,154**	95,958	96,453	97,816	98,18	**98,244**
Correct tokens	22913/ 23402	**22970**/ 23402	22456/ 23402	22572/ 23402	22891/ 23402	22976/ 23402	**22991**/ 23402

Προβληματικές περιπτώσεις: λέξεις

			FBT			MBT			ME			TnT		
word	inst.	tags	class	en_o	%	class	en_o	%	class	en_o	%	class	en_o	%
του	440	3	1	23	5.2	2	25	5.7	2	17	3.9	2	23	5.2
της	408	3	4	17	4.2	9	11	2.7	4	12	2.9	7	12	2.9
τους	125	3	19	5	4	3	19	15.2	8	10	8	17	6	4.8
μία	16	3	10	10	62.5	13	10	62.5	10	10	62.5	12	10	62.5
πριν	10	3	21	3	30	21	3	30	21	3	30	21	3	30

			Majority			Weighted			Combiner			Max. Errors	Errors
word	inst.	tags	class	en_o	%	class	en_o	%	class	en_o	%		
του	440	3	6	13	3	6	12	2.7	10	9	2	17	5
της	408	3	4	13	3.2	5	12	3	8	9	2.2	11	2
τους	125	3	12	7	5.6	16	5	4	14	5	4	5	0
μία	16	3	10	10	62.5	10	10	62.5	7	10	62.5	10	10
πριν	10	3	21	3	30	21	3	30	21	3	30	3	3

Προβληματικές περιπτώσεις: γραμματικές κατηγορίες

		FBT		MBT		ME		TnT	
gold	tagger	class	en_o	class	en_o	class	en_o	class	en_o
AjBa	NoCm	2	30	2	73	2	32	2	57
NoCm	AjBa	1	51	1	97	1	43	1	77
	total		81		170		75		134
NmCd	AtId	5	24	6	24	3	26	5	28
AtId	NmCd	27	4	33	6	24	4		0
	total		28		30		30		28
AjBa	AdXxBa	12	14	11	17	13	10	9	18
AdXxBa	AjBa	11	17	20	13	12	11	28	7
	total		31		30		22		25

		Majority		Weighted		Combiner		Max. Errors	Errors
gold	tagger	class	en_o	class	en_o	class	en_o		
AjBa	NoCm	2	33	2	27	2	28	30	8
NoCm	AjBa	1	53	1	47	1	35	43	9
	total		86		74		63		
NmCd	AtId	3	26	3	25	3	24	24	23
AtId	NmCd	29	4	26	4	26	4	0	0
	total		30		29		28		
AjBa	AdXxBa	13	12	18	7	12	10	10	7
AdXxBa	AjBa	20	7	14	9	16	7	7	1
	total		19		16		17		

Παραδείγματα

❑ **Ένας- μία- ένα (αριθμητικό ή άρθρο;)**

- Η Αθήνα πρόκειται να συμμετάσχει με ένα αρχηγείο Σώματος Στρατού…
- …η παραχώρηση μιας ελληνικής μηχανοκίνητης…

❑ **Ουσιαστικό ή επίθετο;**

- …πραγματοποιηθούν στο Ναυτικό Μουσείο Οινουσών
- Ιρανοί πολιτικοί κρατούμενοι μιλούν …
- …χρονοντούλαπο της πληροφορικής ιστορίας …
- …όνειρο των προημιτελικών**.**
- …της οικονομικής των επιχειρήσεων…
- Οι υπεύθυνοι της εταιρίας …

Παραδείγματα (2)

- **Αντωνυμία ή άρθρο;**
- …της συνέντευξής του επί των …
- …το γιο του Χρ. Νικ. …
- …οι επιδόσεις της υπολείπονται λίγο…
- …στον Ποινικό του Κώδικα…
- Το αίτημά τους ενίσχυσε και …
- …τελευταία έκδοσή του υποστηρίζει το…
- …η δράση τους απειλεί αθώες …

- **Επίθετο ή επίρρημα;**
- …δε σημαίνει απαραίτητα και ότι…
- …να εκκρεμούν ανοιχτά τα μέτωπα…
- …αποδειχτούν αποτελεσματικά σε ορισμένες …

- **Ουσιαστικό ή επίρρημα;**
- Παρόμοια , τέλος συνέβησαν και…
- …για μέρες συνέχεια , περιμένοντας …

Προβλήματα

- Ασυνέπεια στο Σώμα Εκπαίδευσης
- Ανεπάρκεια περικειμένου
- Εξειδίκευση της βασικής κατηγορίας
- Ανεπάρκεια Σώματος Ελέγχου

Προτάσεις

- Αξιόπιστο Σώμα Εκπαίδευσης
- Κοινή χρήση πόρων από base taggers
  - Εφάμιλλα αποτελέσματα
- Active learning
- Επάρκεια Σώματος Ελέγχου
- Επίλυση της αμφισημίας σε μεταγενέστερο επίπεδο
- Μελέτη εναλλακτικών συνδυαστικών μεθόδων
  - Αξιοποίηση πιθανοτήτων των base taggers
- Αξιοποίηση συνδυαστικών μεθόδων στην πλήρη έκταση των μορφολογικών χαρακτηριστικών

References

1. Berger A, Della Pietra S. and Della Pietra V. 1996: *Maximum Entropy Approach to Natural Language Processing.* Computational Linguistics, 22(1).
2. Brants T. 2000: *TnT – A Statistical Part-of-Speech Tagger.* In Proceedings of the 6th Applied Natural Language Processing Conference (ANLP 2000), April 29 – May 3, 2000, Seattle, WA.
3. Brill E. 1995a: *Transformation-Based Error-Driven Learning and Natural Language Processing: a Case Study in Part-of-Speech Tagging*. Computational Linguistics 21(4).
4. Brill E. and Wu J. 1998: *Classifier Combination for Improved Lexical Disambiguation*. In COLING-ACL'98, pp. 191-195, Montreal, Canada, August 10-14.
5. Γιούλη Β., Προκοπίδης Π., Παπαγεωργίου Χ. 2000: *Μορφοσυντακτικός Σχολιασμός κειμένων: Οδηγίες*. Κείμενο Εργασίας, Τμήμα Γλωσσικών Εφαρμογών Γραφείου, ΙΕΛ.
6. Daelemans W., Zavrel J., Berck P. and Gillis S. 1996: *MBT: A memory-based part of speech tagger generator.* In Proceedings of the Fourth Workshop on Very Large Corpora.
7. Daelemans W., Zavrel J., van der Sloot K. and van den Bosch A. 2000: *TiMBL: Tilburg Memory Based Learner, version 3.0, Reference Guide.* Technical Report ILK-0001. ILK, Tilburg University.
8. Dermatas E. and Kokkinakis G. 1995: *Automatic Stochastic Tagging of Natural Language Texts*. Computational Linguistics, 21(2): 137-163.
9. Dietterich T. G. 1997: *Machine Learning Research: Four Current Directions.* AI Magazine, 18(4): 97-136.
10. van Halteren H. (ed.) 1999: *Suntactic Wordclass Tagging.* Dordrecht: Kluwer Academic Publishers.

References (2)

11. van Halteren H., Daelemans W. and Zavrel J. 2001: *Improving Accuracy in Wordclass Tagging through Combination of Machine Learning Systems.* Computational Linguistics 27(2): 199- 230.
12. Koutsombogera M., Konstandinidis A. and Papageorgiou H. 2004: *Combination of Machine Learning Approaches for Error Reduction in POS Tagging*. Proceedings of the 3rd Hellenic Conference on Artificial Intelligence, Samos, 5-8 May 2004. Εκδόσεις Ζήτη, Αθήνα 2004.
13. Labropoulou P., Mantzari E. and Gavrilidou M. 1996: *Lexicon – Morphosyntactic Specifications: Language Specific Instantiation (Greek),* PP-PAROLE, MLAP 63-386 report.
14. Orphanos, G., Christodoulakis, D. 1999: *POS Disambiguation and Unknown Word Guessing with Decision Trees*. In: Proc. EACL 1999, 134-141.
15. Orphanos G., Kalles D., Papagelis T., Christodoulakis D. 1999: *Decision Trees and NLP: A Case Study in POS Tagging.* In proceedings of ACAI'99.
16. Παπαγεωργίου Χ., Πιπερίδης Σ. 1995: *Στατιστική, Κανόνες ή στατιστικοί κανόνες για μορφολογικό σχολιασμό.* Μελέτες για την Ελληνική Γλώσσα, Πρακτικά της 16ης συνάντησης του Τομέα Γλωσσολογίας της Φιλοσοφικής Σχολής του ΑΠΘ, 4- 6 Μαίου 1995.
17. Papageorgiou H., Prokopidis P., Giouli V., Piperidis S. 2000: *A Unified POS Tagging Architecture and its Application to Greek.* In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000). Athens: ELR
18. Petasis, G., Paliouras, G., Karkaletsis, V., Spyropoulos, C. D., Androutsopoulos, I. 1999: *Resolving Part-of-Speech Ambiguity in the Greek Language Using Learning Techniques*. In Proceedings of the ECCAI Advanced Course on Artificial Intelligence (ACAI '99), Chania, Greece.
19. Ratnaparkhi A. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution.* Ph.D. thesis, Department of Computer and Information Science, University of Pennsylvania.

Tagset (basic pos)

- NoCm=common noun, NoPr=proper noun, AjBa=adjective (base form), AjCp=adjective (comparative form), AjSu=adjective (superlative form), AdBa=adverb (base form), AdCp=adverb (comparative form), AdSu=adverb (superlative form), VbMn=main verb or participle, VbIs=impersonal verb, AtDf=definite article, AtId=indefinite article, PnPo=possessive pronoun, PnPe=personal pronoun, PnDm=demonstrative pronoun, PnId=indefinite pronoun, PnIr=interrogative pronoun, PnRe=relative pronoun, PnRi=relative indefinite pronoun, NmCd=cardinal numeral, NmOd=ordinal numeral, NmMl=multiplicative numeral, NmAn=analog numeral, NmCt=collective numeral, AsPpSp=simple preposition, AsPpPa=prepart preposition, CjCo=coordinate conjunction, CjSb=subordinate conjunction, PtFu=future particle, PtNg=negative particle, PtSj=subjunctive particle, PtOt=other particles, RgFwOr=foreign word, RgFwTr=transliterated word, RgAn=acronym, RgSy=symbol, Ij=interjection, DATE=date, DIG=digit, PUNCT=punctuation, ABBR=abbreviation, NBABBR=non-breaking abbreviation, ENUM=enumeration, INIT=initials, TE=tagging error.