

ΥΠΟΛΟΓΙΣΤΙΚΕΣ
ΕΠΙΣΤΗΜΕΣ

ΕΥΦΥΗ
ΒΙΟΜΗΧΑΝΙΚΑ
ΣΥΣΤΗΜΑΤΑ



VIRTUAL EVENT

Ημέρες Δράσης

23-24-25
ΙΟΥΝΙΟΥ
2021



ΤΕΧΝΟΛΟΓΙΕΣ
ΔΕΔΟΜΕΝΩΝ
ΚΑΙ ΑΛΛΗΛΕΠΙΔΡΑΣΗΣ

Έρευνα & Καινοτομία
Τεχνολογίες Πληροφορίας

ΑΘΗΝΑ



ATHENA

Research & Innovation
Information Technologies

Τα παρουσιαζόμενα έργα υλοποιήθηκαν στο πλαίσιο της Πρόσκλησης με κωδικό 031 με τίτλο «ΔΡΑΣΗ ΣΤΡΑΤΗΓΙΚΗΣ ΑΝΑΠΤΥΞΗΣ ΕΡΕΥΝΗΤΙΚΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΚΩΝ ΦΟΡΕΩΝ», που χρηματοδοτείται από το Επιχειρησιακό Πρόγραμμα «Ανταγωνιστικότητα, Επιχειρηματικότητα και Καινοτομία», ΕΣΠΑ 2014-2020 ΕΣΠΑ 2014-2020, με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης (Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης). Αναλυτικά: "Υπολογιστικές Επιστήμες και Τεχνολογίες Δεδομένων, Περιεχομένου και Αλληλεπίδρασης" (MIS 5002437) με 3 υποέργα: Υποέργο 1 «Γλωσσικές Τεχνολογίες Ανάλυσης Περιεχομένου και Αλληλεπίδρασης», Υποέργο 2 «Τεχνολογίες Ανάλυσης Περιεχομένου στον Πολιτισμό», Υποέργο 3 «ΜΕΔΑ: Μεγάλα Δεδομένα – Προκλήσεις, Μέθοδοι και Τεχνικές Αποδοτικής Διαχείρισης» και "Καινοτόμος εφαρμογή του βιομηχανικού διαδικτύου των πραγμάτων σε ευφυή περιβάλλοντα" (MIS 5002434) με Υποέργο 1: "Καινοτόμος εφαρμογή του βιομηχανικού διαδικτύου των πραγμάτων σε ευφυή περιβάλλοντα".



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο
Περιφερειακής Ανάπτυξης



ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ
ΑΝΑΠΤΥΞΗΣ ΚΑΙ ΕΠΕΝΔΥΣΕΩΝ
ΕΙΔΙΚΗ ΓΡΑΜΜΑΤΕΙΑ ΔΙΑΡΘΡΩΤΙΚΩΝ ΠΡΟΓΡΑΜΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ ΕΠΑΝΕΚ

ΕΠΑΝΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ
ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ
ΚΑΙΝΟΤΟΜΙΑ



Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης



ΑΧΑ

Έρευνα & Καινοτομία
Τεχνολογίες Πληροφορίας

ΥΠΟΛΟΓΙΣΤΙΚΕΣ
ΕΠΙΣΤΗΜΕΣ

ΕΥΦΥΗ
ΒΙΟΜΗΧΑΝΙΚΑ
ΣΥΣΤΗΜΑΤΑ



VIRTUAL EVENT

Ημέρες
Δράσης

23-24-25
ΙΟΥΝΙΟΥ
2021



ΤΕΧΝΟΛΟΓΙΕΣ
ΔΕΔΟΜΕΝΩΝ
ΚΑΙ ΑΛΛΗΛΕΠΙΔΡΑΣΗΣ

Εργασία Ε1.5: Βελτίωση Μηχανικής Μετάφρασης

Γ. Ταμπουρατζής

ΙΕΛ/Ε.Κ.Αθηνά



ΑΘΗΝΑ

Έρευνα & Καινοτομία
Τεχνολογίες Πληροφορίας



Στόχος Ενότητας Εργασίας 1.5

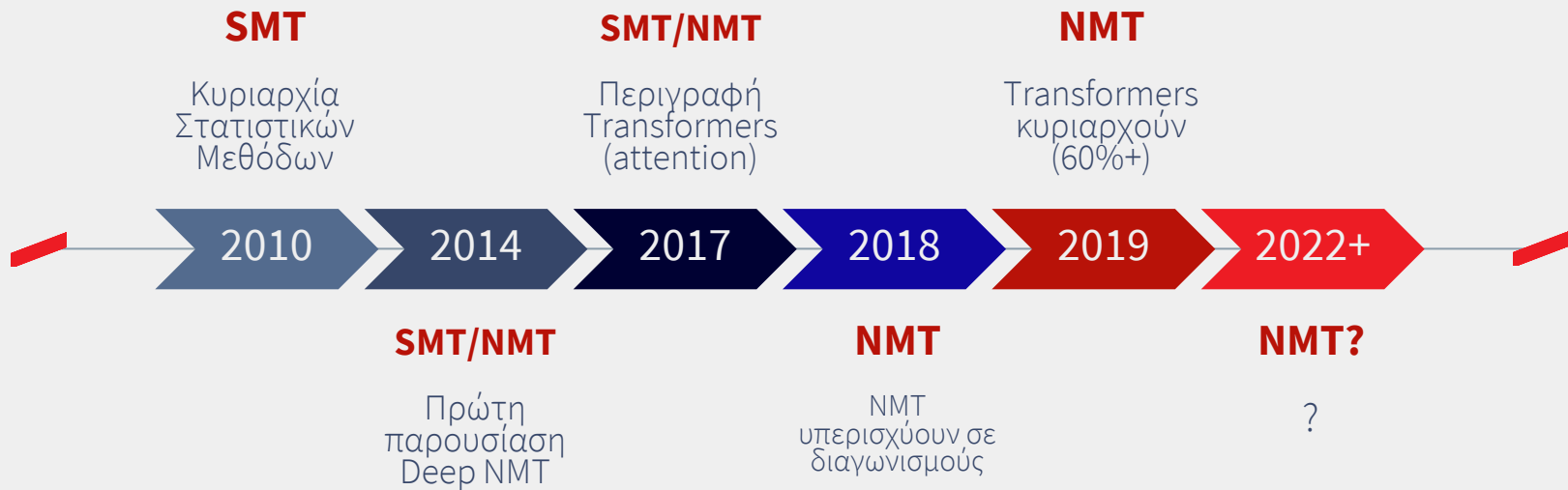


Βελτίωση της μεθοδολογίας δημιουργίας συστημάτων μηχανικής μετάφρασης (MM), η οποία αναπτύχθηκε στο έργο PRESEMT.

Έμφαση δίνεται στη βελτίωση της μετάφρασης για ζεύγη γλωσσών που περιλαμβάνουν την Ελληνική (λ.χ. $EL \rightarrow EN$, $EN \rightarrow EL$ & $EL \rightarrow DE$).



Εξέλιξη Μηχανικής Μετάφρασης





Αντικείμενα Μελέτης NMT

- Βελτιστοποίηση αρχιτεκτονικών NMT (2015-2017)
 - Προσδιορισμός Εγγενών προβλημάτων στη μετάφραση που παράγουν τα συστήματα NMT (2017-2019)
 - Απαιτήσεις συστημάτων NMT – χρειάζονται πολύ περισσότερα δεδομένα (παράλληλα κείμενα) απ' ό,τι τα SMT συστήματα.
- Η λύση τύπου one-fits-all δεν είναι πάντα η καλύτερη. Η προσθήκη πληροφορίας για τη γλώσσα (2018-2020) βελτιώνει την ποιότητα.





Ειδικευμένα θέματα NMT

- Μέθοδοι προεπεξεργασίας κειμένων (2016-2018)
- Ειδική μελέτη για γλώσσες με πλούσια μορφολογία (2017-2019)
- Συλλογή πρόσθετων δεδομένων και χρήση μη αυστηρά παράλληλων δεδομένων (2017+).
- Δραστηριότητες εστιασμένες σε γλώσσες με πολύ λίγα δεδομένα (λ.χ. Lorelei).

Το ζήτημα της αξιόπιστης μετάφρασης έχει καλύτερη λύση πια, αλλά ΔΕΝ έχει λυθεί οριστικά. Η «ιδανική» λύση αργεί, ιδίως για λιγότερο διαδεδομένες γλώσσες (λιγότερα δεδομένα).





ΕΕ.1.4 : 5 Κύριες Κατευθύνσεις Έρευνας

1. Βελτίωση του συστήματος PRESEMT
2. Μελέτη του αρθρώματος PAM, με μεθόδους θεωρίας σμηνών (Particle Swarm Optimisation)
3. Μελέτη κυριαρχούσας μεθοδολογίας - NMT
4. Ανάπτυξη υβριδικών μεθόδων μετάφρασης (NMT+PRESEMT)
5. Προσαρμογή NMT στα χαρακτηριστικά της Ελληνικής

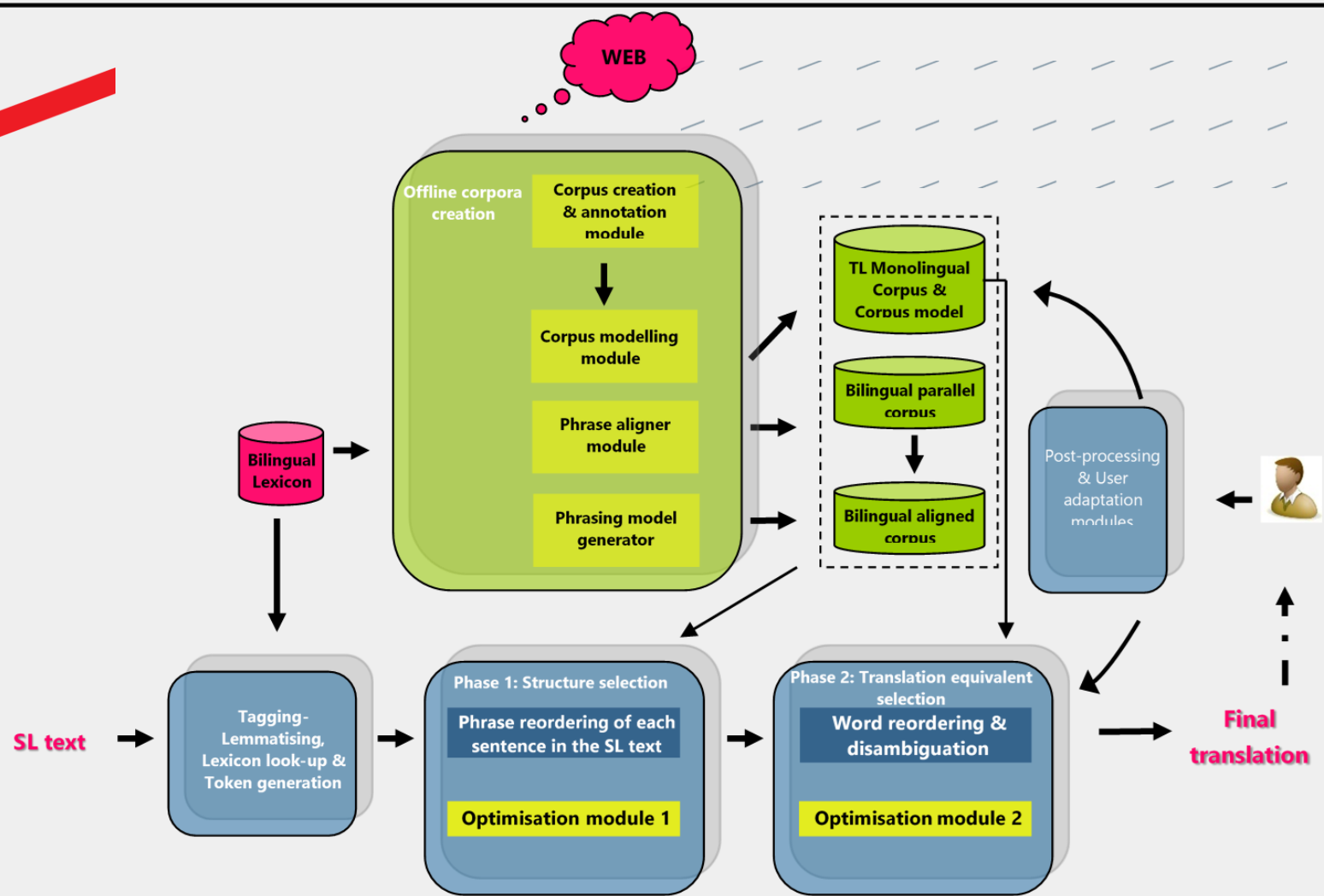




Μεθοδολογία PRESEMT

1. **Βαθμίδα 1: Pre-processing stage (online):** επεξεργασία (επισημείωση με γραμματική & συντακτική πληροφορία) των δεδομένων εισόδου
2. **Βαθμίδα 2: Main translation engine (online):** η βασική διαδικασία μετάφρασης
3. **Βαθμίδα 3: Post-processing stage (offline):** τροποποιήσεις του χρήστη στο μεταφραστικό αποτέλεσμα







ΕΕ.1.4 : Κατεύθυνση (1)

Επαύξηση δυνατοτήτων διαφόρων αρθρωμάτων (modules) του συστήματος PRESEMT

Βελτίωση γλωσσικών πόρων του PRESEMT, με στόχο τον εμπλουτισμό των μοντέλων ώστε να στηρίζεται η επακριβής αναζήτηση πληροφορίας.

Βελτιστοποίηση των παραμέτρων του συστήματος με συστηματικές μεθόδους υπολογιστικής νοημοσύνης (swarm intelligence)





Αποτελέσματα – Κατεύθυνση 1

Μέτρηση ακρίβειας ποιότητας μετάφρασης (EN-to-EL)

Έμφαση στη μετάφραση προς τη μορφολογικά πλούσια γλώσσα

Εκτίμηση ποιότητας με τις μετρικές BLEU/NIST

	BLEU	NIST
1 ^η φάση	0,1014	3,4483
2 ^η φάση	0,1294	3,9477





ΕΕ.1.4 : Κατεύθυνση (2)

Έμφαση δόθηκε στο υποσύστημα διαχωρισμού φράσεων του PRESENT σε παράλληλο κείμενο

Βελτιστοποίηση των βαρών (χρήση μοντέλου ελκτικών-απωστικών δυνάμεων)

Χρήση μεθόδων υπολογιστικής νοημοσύνης (swarm intelligence) σε απλοποιημένες συναρτήσεις (surrogate functions) με μειωμένο αριθμό παραμέτρων, για καλύτερη βελτιστοποίηση



Αποτελέσματα – Κατεύθυνση 2





ΕΕ.1.4 : Κατεύθυνση (3)

Μελέτη νέων μεθοδολογιών μετάφρασης.

- Μελέτη συστημάτων βασισμένων σε Πολυεπίπεδα Νευρωνικά δίκτυα (υιοθέτηση MarianMT)
- Συλλογή πόρων (παράλληλων κειμένων) και υλοποίηση συστημάτων NMT
- Εκπαίδευση 3 διαφορετικών μοντέλων (transformer, s2s, amun)



Αποτελέσματα – Κατεύθυνση 3

Model	Dataset1		Dataset2		Improvement over PRESENT	
	BLEU	NIST	BLEU	NIST	BLEU	NIST
Transformer	0.3496	6.2282	0.1884	4.2543	+45.6%	+7.8%
Amun	0.3777	6.5240	0.1883	4.1289	+45.5%	+4.6%
S2S	0.3558	6.4756	0.2048	4.5263	+58.3%	+14.7%





#1 Input: Obviously, he wants to downplay them, and tell us that whatever the people do, the American, European, Japanese monopolies and the rest will impose their will and overcome these reactions together.

#1 Translation: ωστόσο , πρέπει να διαδραματίσει επίσης τον ρόλο του , και να μας πει ότι , όταν οι άνθρωποι κάνουν , οι αμερικανοί , οι ευρωπαίοι , οι Αμερικανοί και οι Αμερικανοί θα έχουν τη βούλησή τους και θα έχουν την πλήρη υποστήριξή τους .





#2 Input: In reality, the problems and contradictions of the global economy are still there, unresolved and still relevant: doubts about genetically modified organisms, child labour, the negative social effects of unbridled, unregulated globalisation.

#2 Translation: επιπλέον , οι στόχοι και οι στόχοι της παγκόσμιας οικονομίας είναι σημαντικοί εκεί, η απασχόληση και η κοινωνική συνοχή : η καταπολέμηση της φτώχειας , η καταπολέμηση της φτώχειας , η καταπολέμηση της παιδικής φτώχειας , η καταπολέμηση της φτώχειας , η καταπολέμηση της φτώχειας .





ΕΕ.1.4 : Κατεύθυνση (4)

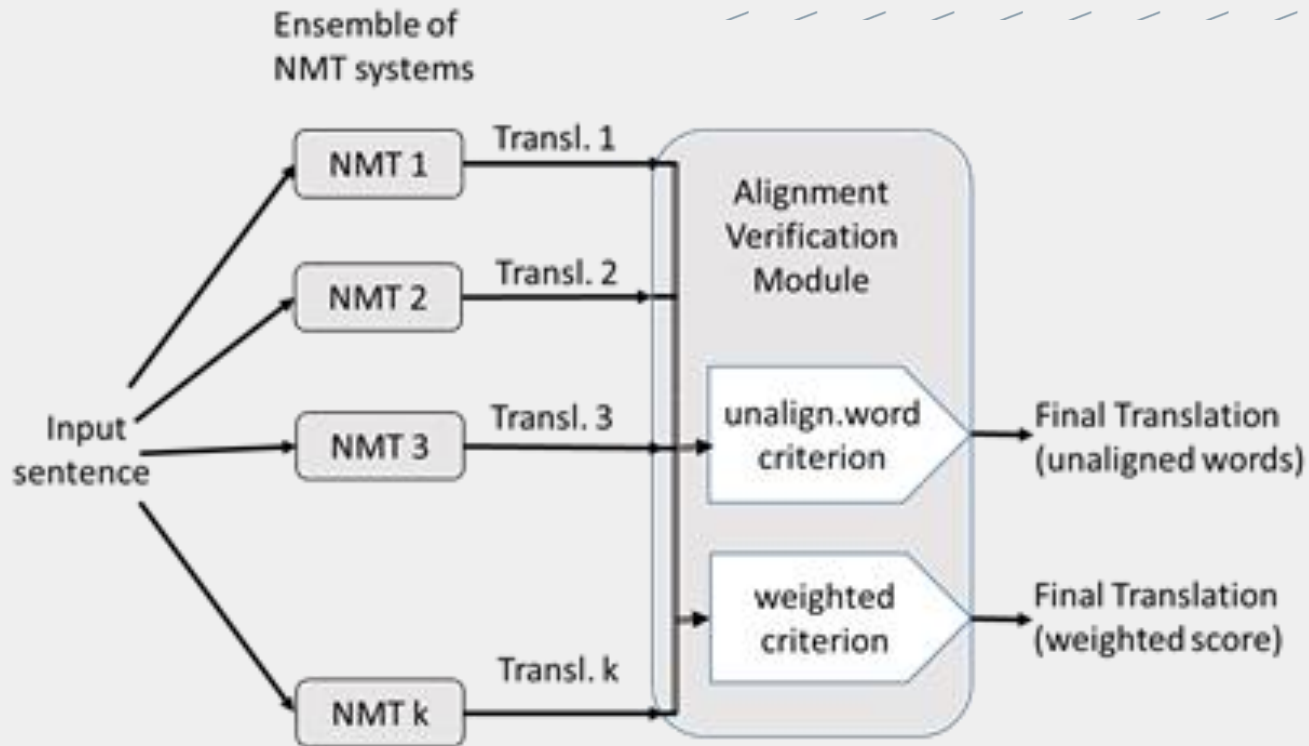
Υβριδοποίηση NMT (χρήση αρθρωμάτων του PRESEMT σε συνδυασμό με NMT αρχιτεκτονικές) με στόχο τη βελτίωση της μετάφρασης

Στόχος είναι ο βέλτιστος συνδυασμός συστημάτων με όμοιες ή και διαφορετικές αρχιτεκτονικές (ensemble).

Η μέθοδος βασίζεται στην εκτίμηση του βαθμού παραλληλίας μεταξύ πρότασης εισόδου και μετάφρασης για την επιλογή του καλύτερου μεταφραστικού συστήματος.

Ορίσθηκαν οι μετρικές WScore (weighted) & UScore (unweighted) για τον προσδιορισμό της πιστότητας μετάφρασης.





Αποτελέσματα – Κατεύθυνση 4

Testset 1 / Transformer

method	Criterion	BLEU	NIST
Transf. NMT	BLEU	0.3522	6.2377
	Entropy	0.3480	6.2132
	Final	0.3496	6.2224
	Marian-ensemble	0.3563	6.3049
PAM + transf	Uscore	0.3394	6.1277
	Wscore	0.3599*	6.3840*

Testset 2 / Transformer

method	Criterion	BLEU	NIST
Transf. NMT	BLEU	0.1827	4.0996
	Entropy	0.2041	4.9435
	final	0.1884	4.2543
	Marian-ensemble	0.2648	6.4538
PAM + marian	Uscore	0.2635	6.4070
	Wscore	0.2668*	6.3632





Sent id	Source text	Marian-ensemble translation	Wscore-ensemble translation
648	That amendment <u>establishes three or four very important changes: firstly, the Canary Islands were incorporated into the Common agricultural policy and the Common fisheries policy; secondly, they were incorporated into the Community Customs Union, and are therefore subject to ordinary customs duties; and thirdly, .. to protect local industry.</u>	Η τροποποίηση του κανονισμού (ΕΟΚ) αριθ.	Η τροπολογία για τη σύσταση τριών ή τεσσάρων πολύ σημαντικών αλλαγών: Πρώτον, οι Καναρίους Νήσους ενσωματώθηκαν στην κοινή γεωργική πολιτική και στην κοινή αλιευτική πολιτική · δεύτερον, ενσωματώθηκαν στην τελωνειακή ένωση της Κοινότητας και, συνεπώς, υπόκεινται σε συνήθεις τελωνειακούς δασμούς · και τρίτον, .. την προστασία της τοπικής βιομηχανίας.
774	<u>Thirdly, who gets the money ?</u>	<NULL>	Ποιος λαμβάνει τα χρήματα;
962	Furthermore, we have been abolishing subsidies in the shipbuilding sector, <u>yet it has continued to fall apart.</u>	Ωστόσο, καταργήσαμε τις επιδοτήσεις στον τομέα της ναυπηγικής βιομηχανίας.	Για παράδειγμα, καταργήσαμε τις επιδοτήσεις στον ναυπηγικό τομέα, αλλά εξακολουθεί να υφίσταται.





ΕΕ.1.4 : Κατεύθυνση (5)

Ενσωμάτωση πληροφορίας σχετικής με τη μορφολογία της Ελληνικής γλώσσας για βελτίωση της ακρίβειας της μετάφρασης

Χρήση μεθόδων βασισμένων σε στατιστική πληροφορία (AMP – Automatic Morphological Processor) για προεπεξεργασία των δεδομένων εκπαίδευσης (παράλληλων κειμένων)

Συγκριτική μελέτη συστημάτων για προσδιορισμό της βελτίωσης με εισαγωγή μορφολογικής πληροφορίας



Αποτελέσματα – Κατεύθυνση 5

		Testset 1			Testset2		
criterion	normalisation	BLEU	NIST	# null transl.	BLEU	NIST	# null transl.
(No amp)	0.6	0.2827	6.6987	13	0.3957	6.7110	5
Shortest-end	0.6	0.2556	6.2871	44	0.4188	6.9446	3
Minmax	0.6	0.2333	5.9872	84	0.3934	6.8498	7
Shortest-end	1.5	0.2675	6.2602	0	0.4020	6.7060	0
Minmax	1.5	0.2629	6.1607	0	0.3881	6.5942	0





VIRTUAL EVENT

Ημέρες Δράσης

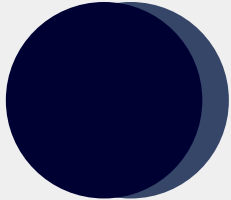
23-24-25
ΙΟΥΝΙΟΥ
2021



Thank you!

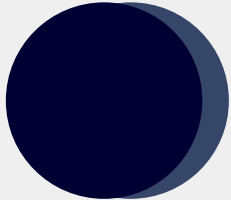


Συμμετέχοντες



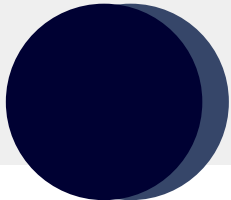
Μ. Βασιλείου

Ε.Κ.Αθηνά



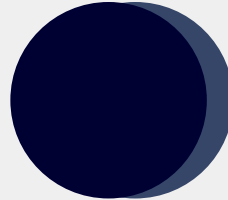
Ι. Κουλιεράκης

Ε.Κ.Αθηνά



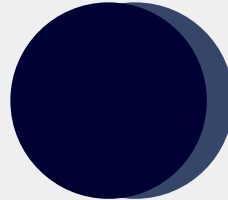
Σ. Μπακαμίδης

Ε.Κ.Αθηνά



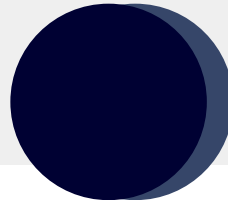
Θ.Γουλάς

Ε.Κ.Αθηνά



Ι. Δολόγλου

Ε.Κ.Αθηνά



Γ. Ταμπουρατζής

Ε.Κ.Αθηνά

