POS Tagging for CS Data

EMNLP 2016

Fahad AlGhamdi, Mona Diab, Abdelati Hawari
The George Washington University
Giovanni Molina, Thamar Solorio
University of Houston
Victor Soto, Julia Hirschberg
Columbia University

Outline

- Introduction
  - Motivation.
  - Main Contribution.
- Approach
- Evaluation
- Discussion
- Conclusion

Introduction

- **Code Switching:** Linguistic Code Switching (CS) is a phenomenon that occurs when multilingual speakers alternate between two or more languages or dialects.
- **Example:**
  - **Arabic Intra-sentential CS**: wlkn AjhztnA AljnAgyp lAnhA m$ xyAl Elmy lm tjd wlw mElwmp wAHdp.
  - **English Translation:** Since our crime investigation departments are not dealing with science fiction, they did not find a single piece of information.

Introduction: Motivation

- Addressing the problem of part of Speech tagging (POS) for CS data on the intra-sentential level.

- Focusing on two language pairs Spanish-English (SPA-ENG) and Modern Standard Arabic and the Egyptian Arabic dialect (MSA-EGY).

- Using the same POS tag sets for both language pairs, the Universal POS tag set (Petrov et al., 2011)

Introduction: Our Contribution

- Exploring different strategies to leverage monolingual resources for POS tagging CS data.

- Presenting the first empirical evaluation on POS tagging with two different language pairs.

- All of the previous work focused on a single language pair combination.

Outline

- Introduction
  - ✓Motivation.
  - ✓Main Contribution.
- Approach
  - Monolingual POS Tagging systems
  - Combined Experimental Conditions.
  - Integrated Experimental Conditions.
- Evaluation
- Discussion
- Conclusion

Approach

- We adopt a supervised framework for our experimental set up.
- we compare leveraging monolingual state of the art POS taggers using different strategies in what we call **a COMBINED framework** comparing it against using a single CS trained POS tagger identified as an **INTEGRATED framework**.
- We explore different strategies to investigate the optimal way of tackling POS tagging of CS data.

Approach: Monolingual POS Tagging Systems

MSA - EGY Language Pair:

- We used the publicly available MADAMIRA tool.
- It is fast, comprehensive tool for morphological analysis and disambiguation of Arabic.
- MADAMIRA MSA is trained on newswire data (Penn Arabic Treebanks 1,2,3).
- MADAMIRA EGY is trained on Egyptian blog data which comprises a mix of MSA, EGY and CS data (MSA-EGY) from the LDC Egyptian Treebank parts 1-5 (ARZ1-5) (Maamouri et al., 2012).

Approach: Monolingual POS Tagging Systems

MSA - EGY Language Pair:

- we need a relatively pure monolingual tagger per language variety (MSA or EGY) trained on informal genres for both MSA and EGY.
  - we retrained a new version of MADAMIRAMSA strictly on pure MSA sentences identified in the EGY Treebank ARZ1-5.
  - we created a MADAMIRA-EGY tagger trained specifically on the pure EGY sentences extracted from the same ARZ1-5 Treebank.

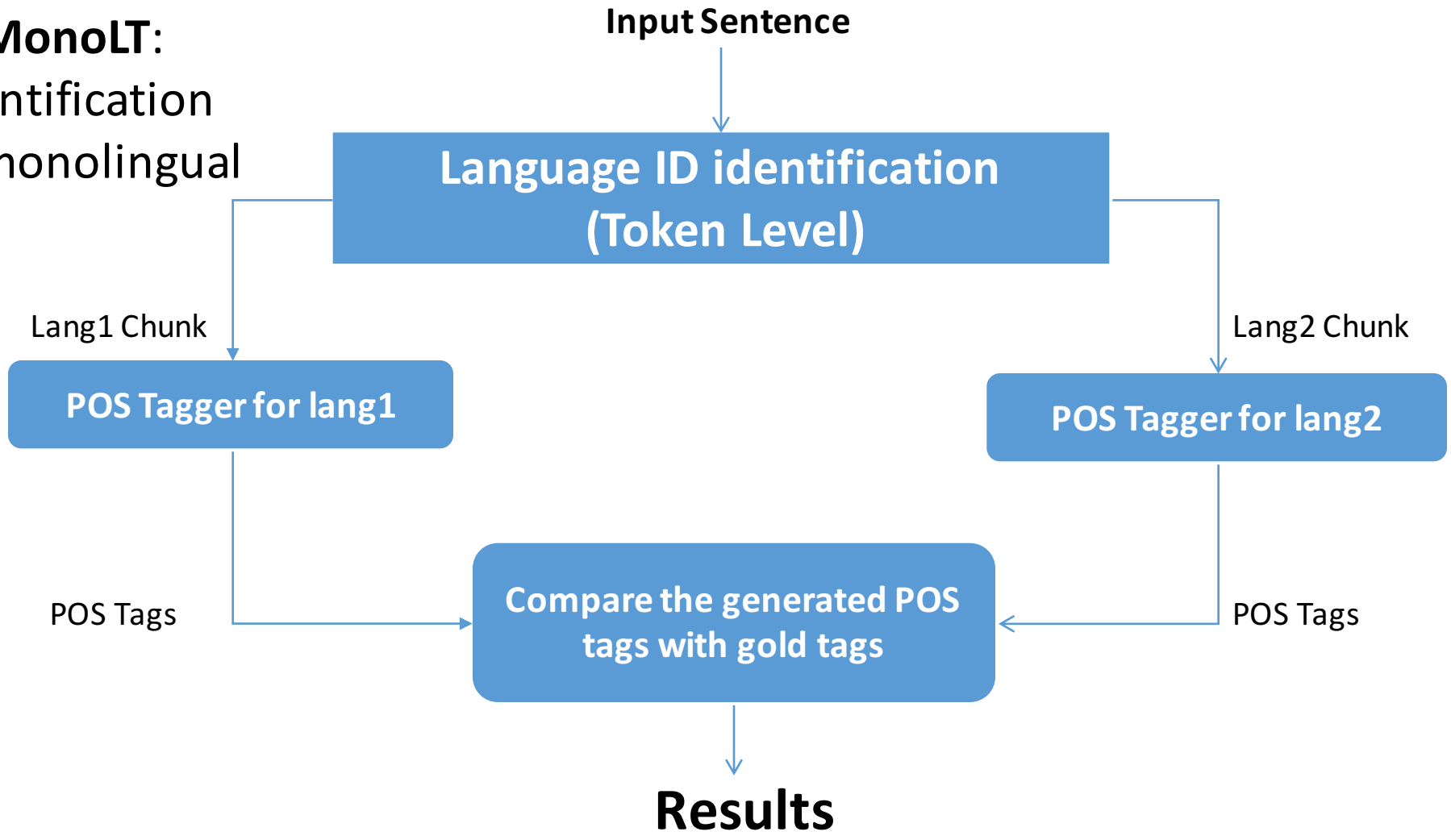
Approach: Monolingual POS Tagging Systems

SPA - ENG Language Pair:

- we created models using the TreeTagger monolingual systems for Spanish and English respectively.
- The data used to train TreeTagger for English was the Penn Treebank data (Marcus et al., 1993), sections 0-22.
- For the Spanish model, we used Ancora-ES.

Approach: Combination Experimental Conditions

COMB1:LID-MonoLT: Language identification followed by monolingual tagging.



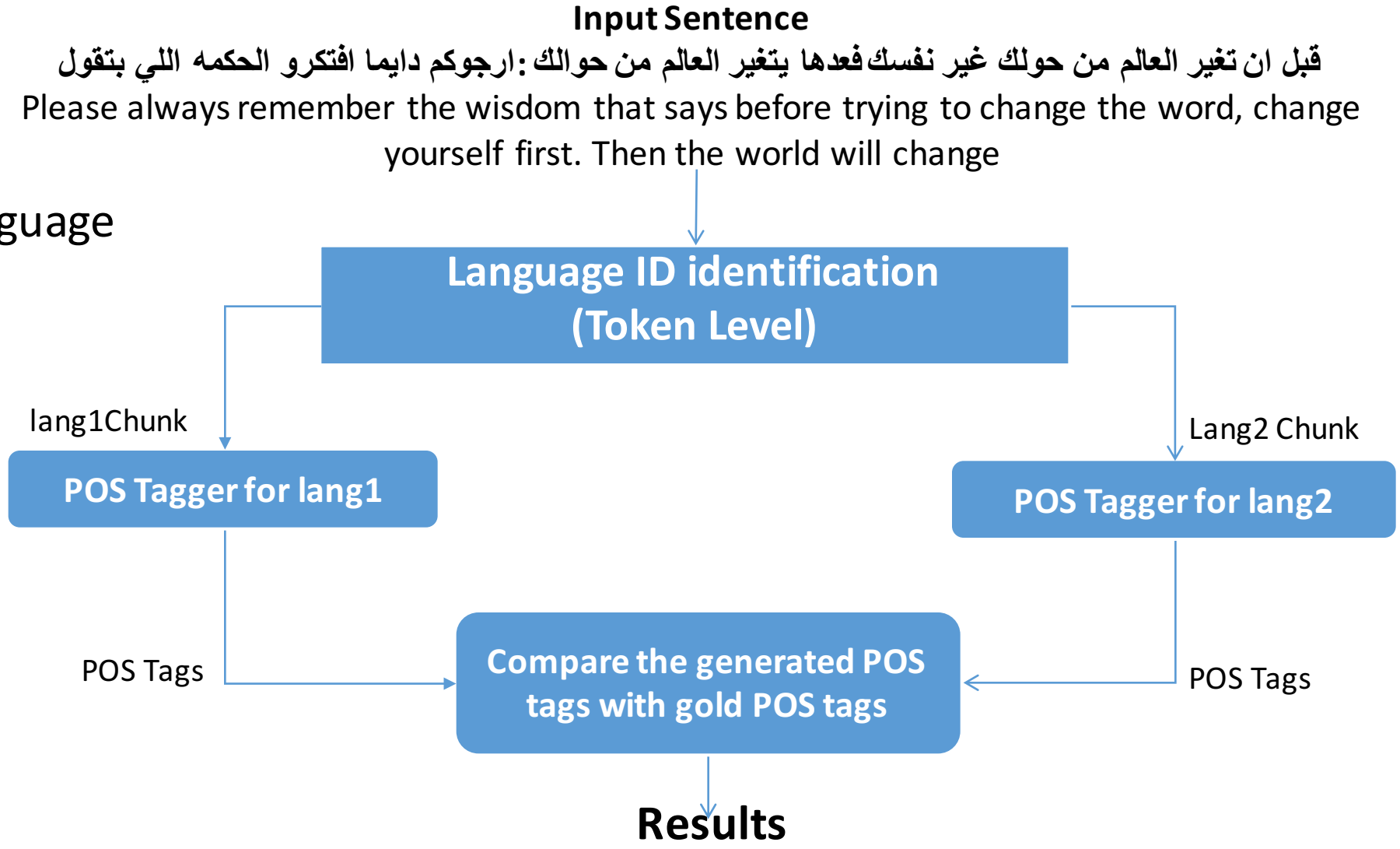
Approach: Combination Experimental Conditions

COMB1:LID-MonoLT:

- **For MSA-EGY:** We used the Automatic Identification of Dialectal Arabic (AIDA2) tool to perform token level language identification for the EGY and MSA tokens in context.
- **For SPA-ENG:** We trained 6-gram character language models using the SRILM Toolkit.
  - The English language model was trained on the AFP section of the English GigaWord.
  - The Spanish language model was trained on the AFP section of the Spanish GigaWord

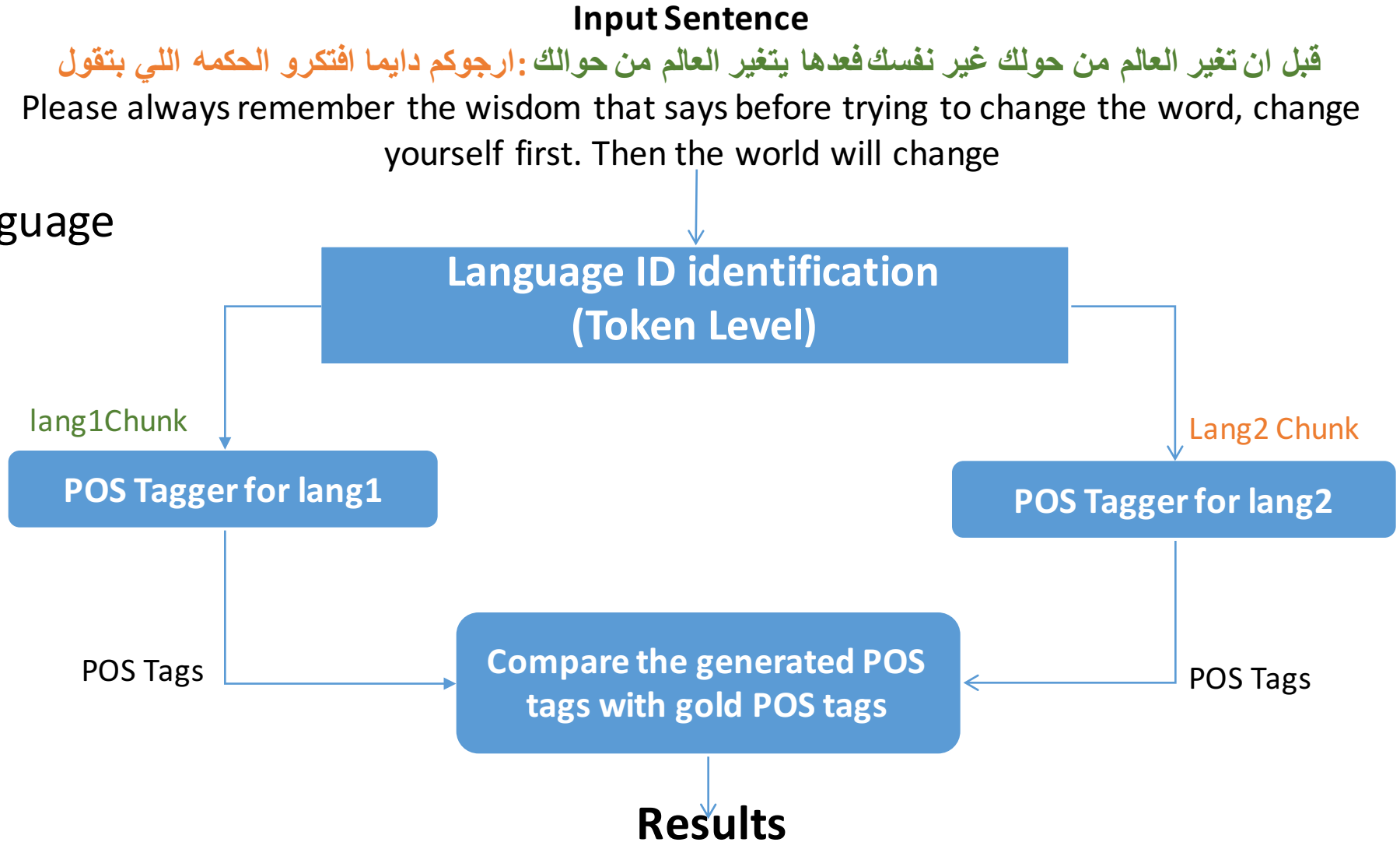
Approach: Combination Experimental Conditions

COMB1:LID-MonoLT: Language identification followed by monolingual tagging.



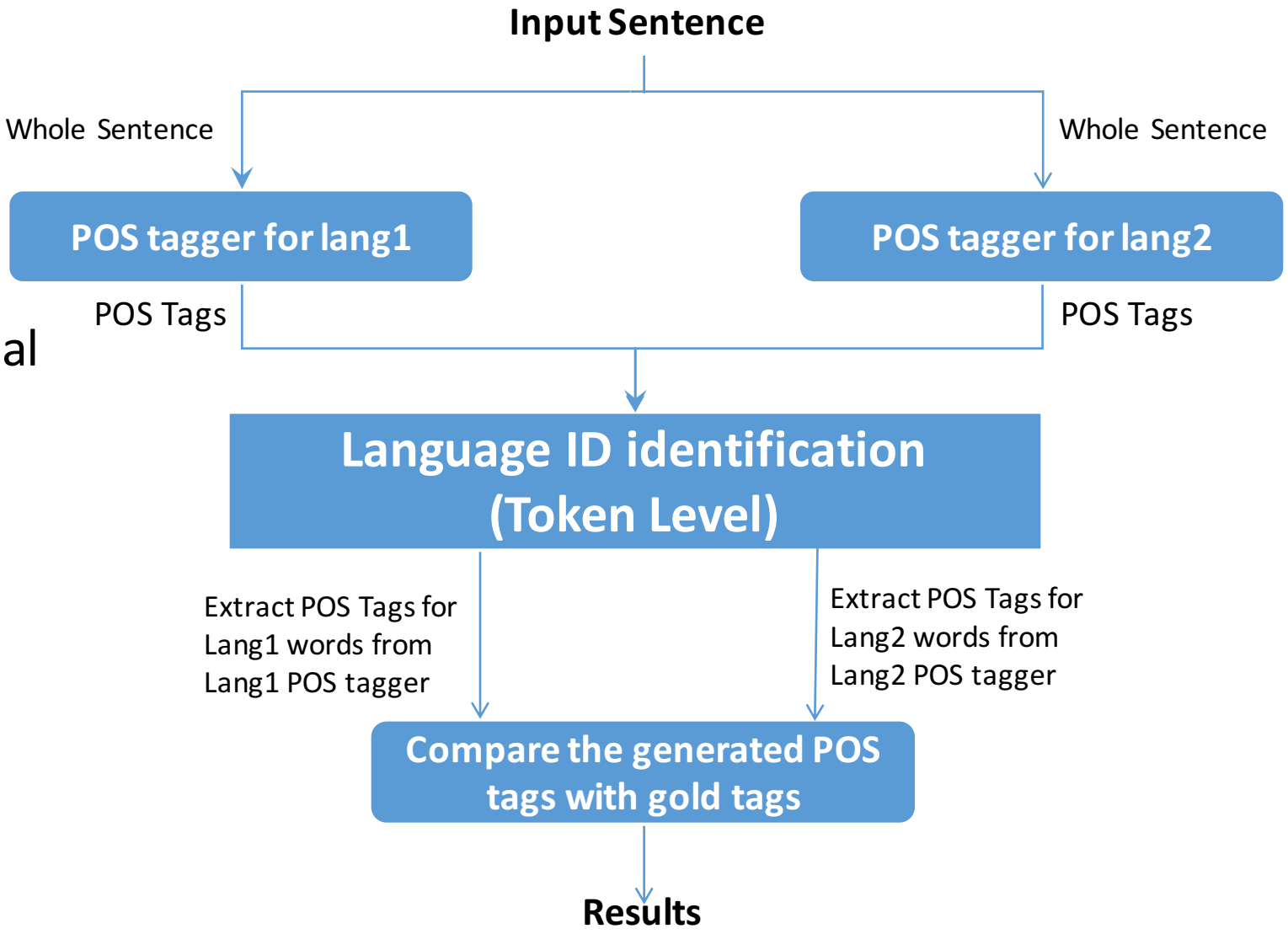
Approach: Combination Experimental Conditions

COMB1:LID-MonoLT: Language identification followed by monolingual tagging.



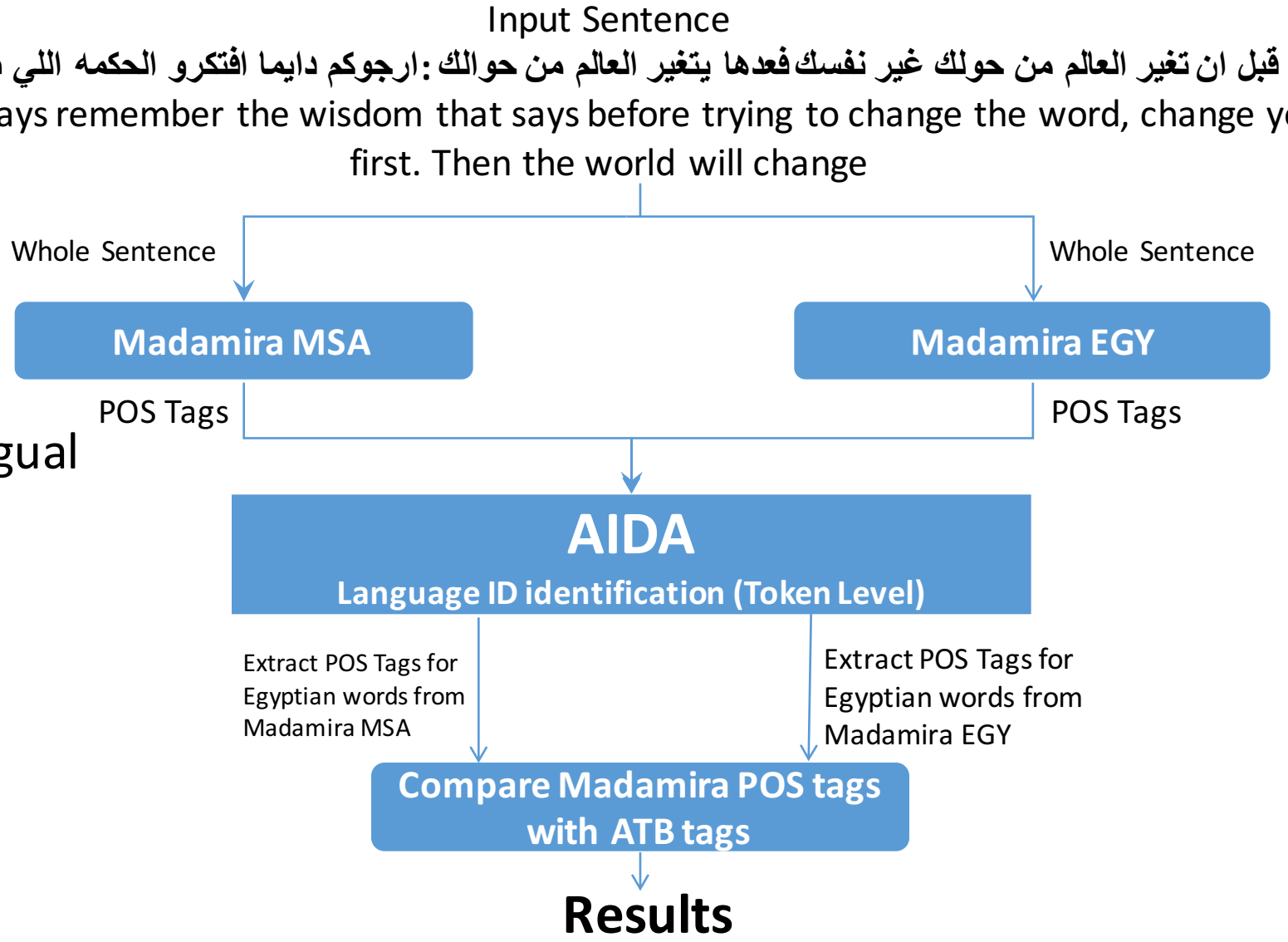
Approach: Combination Experimental Conditions

COMB2:MonoLT-LID: Monolingual tagging then Language ID.



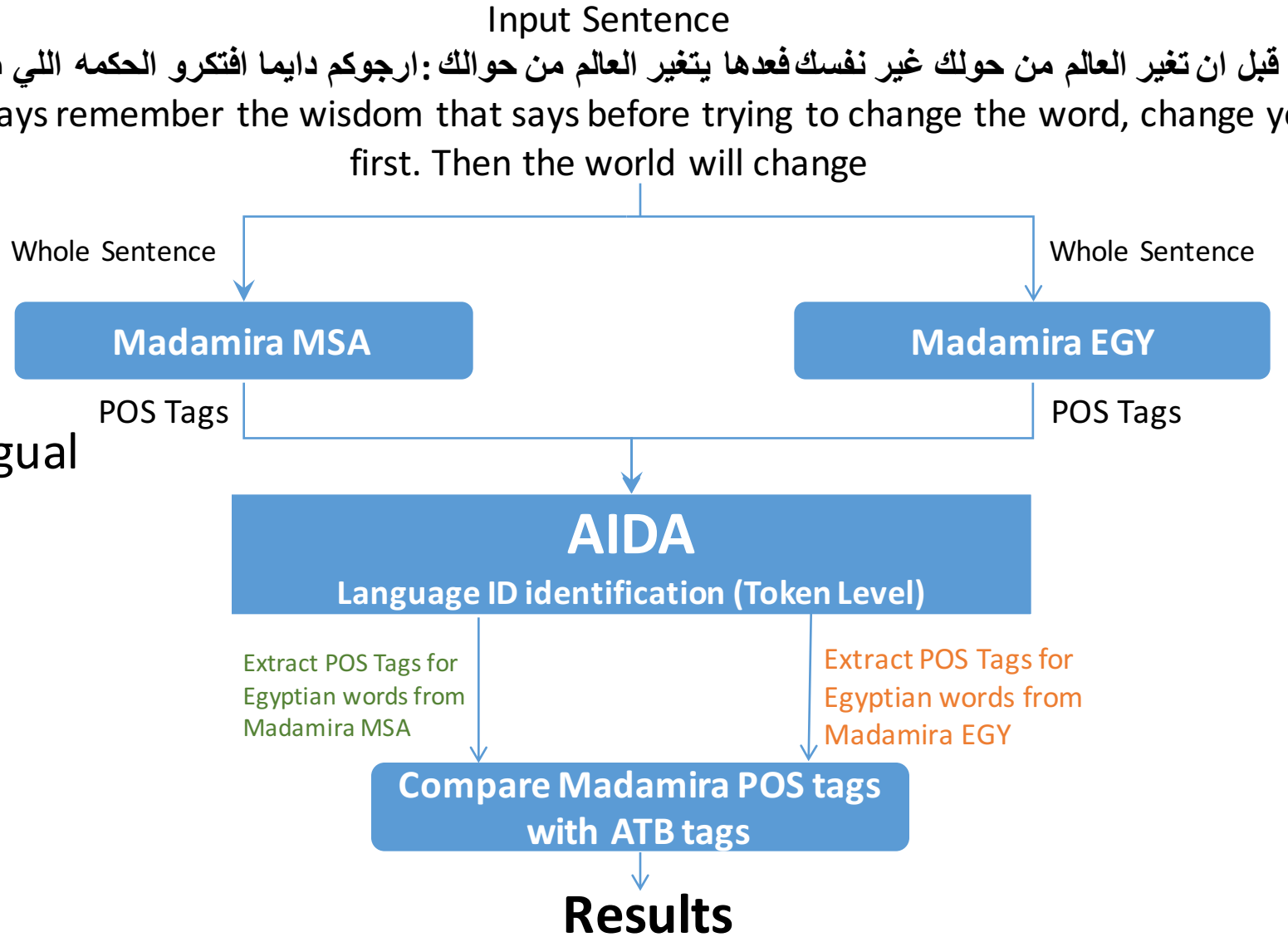
Approach: Combination Experimental Conditions

COMB2:MonoLT-LID: Monolingual tagging then Language ID.



Approach: Combination Experimental Conditions

COMB2:MonoLT-LID: Monolingual tagging then Language ID.



Approach: Combination Experimental Conditions

- **COMB3:MonoLT-Conf:**
  - Apply separate taggers.
  - Then use probability/confidence scores yielded by each tagger to choose which tagger to trust more per token.

- **COMB4:MonoLT-SVM:**
  - Combining results from the monolingual taggers (baselines) and COMB3 into an ML framework such as SVM to decide which tag to choose from (MSA vs. EGY for example or SPA vs. ENG).

Approach: Integrated Experimental Conditions

- **INT1:CSD:** Train a supervised ML framework on exclusively code switched data.
  - **For MSA – EGY:** Train a MADAMIRA model exclusively with the CS data
  - **For SPA – ENG:** Trained a CS model using TreeTagger.

- **INT2:AllMonoData:** Similar to Condition INT1:CSD but changing the training data for each of the language pairs.
  - **For MSA – EGY:** merging the training data from MSA and EGY.
  - **For SPA – ENG:** merging the Spanish and English corpora creating an integrated SPA-ENG model.

Approach: Integrated Experimental Conditions

- **INT3:AllMonoData+CSD:** Merging training data from conditions “INT1:CSD” and ”INT2:AllMonoData” to train new taggers for CS POS tagging.

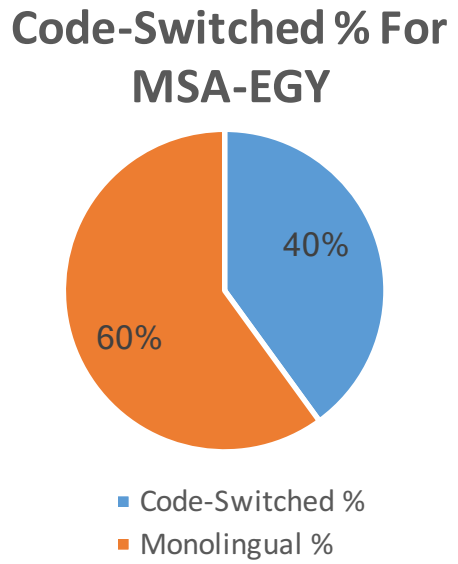
Outline

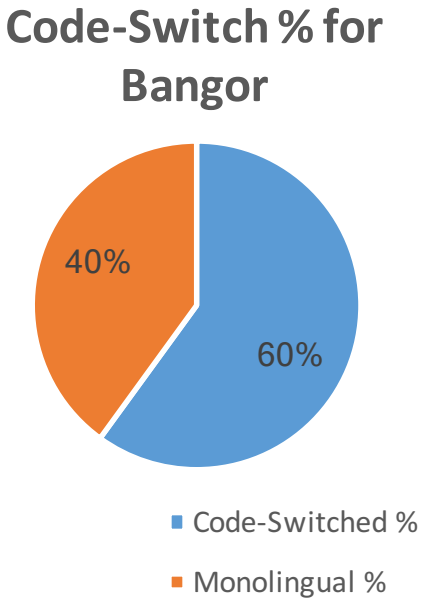
- Introduction
  - ✓Motivation.
  - ✓Main Contribution.
- Approach
  - ✓Monolingual POS Tagging systems
  - ✓Combined Experimental Conditions.
  - ✓Integrated Experimental Conditions.
- Evaluation
  - Datasets
  - POS Tag Sets
  - Results
- Discussion
- Conclusion

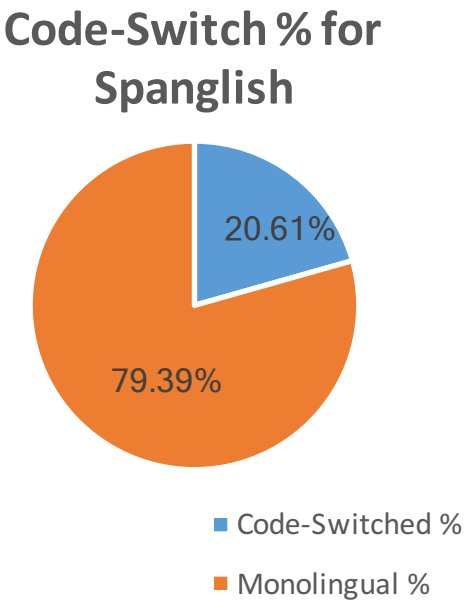
Evaluation: Datasets

- **MSA – EGY:**
  - We use the LDC Egyptian Arabic Treebanks 1-5 (ARZ1-5).
  - The ARZ1-5 data is from the discussion forums genre mostly in the Egyptian Arabic dialect (EGY).

- **SPA – ENG:** Two datasets that were used for SPA-ENG language pair
  - The transcribed conversation used in the work by Solorio and Liu (Solorio and Liu, 2008), referred to as **Spanglish.**
  - **The Bangor Miami corpus**, referred to as Bangor. This corpus is conversational involving a total of 84 speakers living Miami, FL







Evaluation: Datasets

Dataset	# Sentences	# Words	# Types	CS %
ARZ	13,698	175,361	39,168	38.78%
Spanglish	922	8,022	1,455	20.61%
Bangor	45,605	335,578	13,994	6.21%

Table-1: Data set details.

Dataset	Train/Dev Tokens	Test Tokens
ARZ	154,897	20,464
Spanglish	6,456	1,566
Bangor	268,464	67,114

Table-1: Data set distribution.

Evaluation: Part-of-Speech Tag set

MSA-EGY:

- The ARZ1-5 data set is manually annotated using the Buckwalter (BW) POS tag set.
- The BW POS tag set is considered one of the most popular Arabic POS tag sets.

SPA-ENG:

- The Bangor Miami corpus has been automatically glossed and tagged with part-of-speech tags in the following manner:
  - each word is automatically glossed using the Bangor Autoglosser.

Evaluation: Part-of-Speech Tag set

BW POS tag set	Mapping to Universal POS tag set
1- Personal, relative, demonstrative, interrogative, and indefinite pronouns.	Mapped to Pronoun.
2-Acronyms.	Mapped to Proper Nouns.
3- Complementizers and adverbial clause introducers.	Mapped to Subordinating Conjunction.
4- Main verbs (content verbs), copulas, participles, and some verb forms such as gerunds and infinitives.	Mapped to Verb.
5- Prepositions and postpositions.	Mapped to Adpositions.
6- Interrogative, relative and demonstrative adverbs.	Mapped to Adverb.
7- Tense, passive and Modal auxiliaries.	Mapped to Auxiliary Verb.
8- Possessive determiners, demonstrative determiners, interrogative determiners, quantity/ quantifier determiners, etc.	Mapped to Determiner.
9- Noun and gerunds and infinitives.	Mapped to Noun.
10- Negation particle, question particle, sentence modality, and indeclinable aspectual or tense particles	Mapped to Particle

Table: Mapping table for BW POS tag set and and Universal POS tag set

Evaluation: Part-of-Speech Tag set

SPA-ENG:

- The Bangor corpus went through two edition/annotation stages:

1. **Stage one includes:**
   a) Tokens that tagged ambiguously with more than one POS tag were disambiguated (e.g. that .CONJ.[or].DET).
   b) Ambiguous POS categories like ASV, AV and SV were disambiguated into either ADJ, NOUN, or VERB.
   c) For frequent tokens like so and that, their POS tags were hand-corrected.
   d) Mistranscribed terms which were originally labeled as Unknown were hand corrected and given a correct POS tag

Evaluation: Part-of-Speech Tag set

SPA-ENG:

1. **Stage two includes:**
   - Mapping the Bangor corpus original POS tagset to the Universal POS tag set.

Evaluation: Part-of-Speech Tag set

Bangor POS tag set	Mapping to Universal POS tag set
1- Exclamations and Intonational Markers.	Mapped to Interjections
2- Possessive Adjectives, Possessive Determiners, Interrogative Adjectives, Demonstrative Adjectives and Quantifying Adjective	Mapped to Determiner
3- Relatives, Interrogatives and Demonstratives (with no specification to whether they were Determiners, Adjectives or Pronouns).	Manually labeled
4- possessive markers, negation particles, and infinitive to tokens.	PRT
5- Conjunctions	Mapped to Coordinating Conjunctions and Subordinating Conjunctions using word lists
6- A subset of English Verbs	Mapped to Auxiliary Verbs (could, should, might, may, will, shall, etc.
7- Categories with an obvious match (like Nouns, Adjectives, Verbs, Pronouns, Determiners, Proper Nouns, Numbers, etc.)	Automatically mapped to the appropriate category

Table: Mapping table for Bangor POS tag set and and Universal POS tag set

Evaluation: Results

To evaluate the performance of our approaches:

- Comparing the output POS tags generated from each condition against the available gold POS tags for each data set.

- Compare the accuracy of our approaches for each language pair to its corresponding monolingual tagger baseline.

Evaluation: Results

MSA-EGY Baseline		
Data set	**MADAMIRA-MSA**	**MADAMIRA-EGY**
ARZ	77.23 %	72.22 %
SPA-ENG Baseline		
Data set	**TreeTagger SPA**	**TreeTagger ENG**
Spanglish	44.61 %	75.87 %
Bangor	45.95 %	64.05 %

Table: POS tagging accuracy for monolingual baseline taggers

Evaluation: Results

Approach	Overall	CS Posts	MSA Posts	EGY Posts
COMB1:LID-MonoLT	77.66	78.03	76.79	78.57
COMB2:MonoLT-LID	77.41	77.41	78.31	77.01
COMB3:MonoLT-Conf	76.66	77.89	76.79	76.11
COMB4:MonoLT-SVM	**90.56**	**90.85**	**91.63**	**88.91**
INT1:CSD	83.89	82.03	82.48	83.26
INT2:AllMonoData	87.86	87.92	86.82	86
INT3:AllMonoData+CSD	89.36	88.12	85.12	87

Table: Accuracy Results for ARZ Test Data set

Baseline:

MSA:77.23%

EGY: 72.22 %

Evaluation: Results

Approach	Overall	CS Posts	ENG Posts	SPA Posts
COMB1:LID-MonoLT	68.35	71.11	66.36	76.02
COMB2:MonoLT-LID	65.51	69.66	64.44	71.32
COMB3:MonoLT-Conf	68.25	68.21	71.93	65.03
COMB4:MonoLT-SVM	**96.31**	**95.39**	**96.37**	**96.60**
INT1:CSD	95.28	94.41	94.41	95.15
INT2:AllMonoData	78.57	78.62	81.85	76.53
INT3:AllMonoData+CSD	91.04	89.59	92.00	89.48

Baseline:

SPA: 44.61 %

ENG: 75.87 %

Table: Accuracy Results for Bangor Data set

Evaluation: Results

Approach	Overall	CS Posts	ENG Posts	SPA Posts
COMB1:LID-MonoLT	78.73	77.81	80.18	73.99
COMB2:MonoLT-LID	73.52	73.80	73.60	71.57
COMB3:MonoLT-Conf	77.39	76.11	80.20	65.43
COMB4:MonoLT-SVM	**90.61**	**89.43**	**93.61**	**87.96**
INT1:CSD	82.95	83.03	85.95	77.26
INT2:AllMonoData	84.55	84.84	88.50	76.59
INT3:AllMonoData+CSD	85.06	84.70	90.15	76.59

Table: Accuracy Results for Spanglish Data set

Baseline:

SPA:45.95 %

ENG: 64.05 %

Outline

- Introduction
  - ✓Motivation.
  - ✓Main Contribution.
- Approach
  - ✓Monolingual POS Tagging systems
  - ✓Combined Experimental Conditions.
  - ✓Integrated Experimental Conditions.
- Evaluation
  - ✓Datasets
  - ✓POS Tag Sets
  - ✓Results
- Discussion
  - Combined Conditions.
  - Integrated Conditions.
- Conclusion

Discussion: Combined conditions

For MSA – EGY:

- All the combined experimental conditions outperform the baselines
- COMB1:LID-MonoLT yields worse results than COMB2:MonoLT-LID.
- It is expected due to the fact that the taggers are expecting well formed sentences on input.
- The worst results are for condition MonoLT-Conf.

For SPA– ENG: Spanglish data set:

- Almost all the accuracies achieved by the combined conditions are higher than the Spanglish data set's baselines.
- "COMB2:MonoLT-LID" is the only combined condition that is lower than the baselines' of the Spanglish data set (73.52%, 75.87%).

Discussion: Combined conditions

For SPA– ENG: Spanglish data set:

- Mistakes in the automated language identification that causes the wrong tagger to be chosen.

For SPA– ENG: Bangor data set:

- All the accuracies achieved by the combined conditions are higher than the Bangor data set's baselines.

Discussion: Combined conditions

- The trends **almost** the same between the two language pairs.
- Both language pairs achieve the highest performance with MonoLT-SVM and worse results with MonoLT-Conf.
- The weaknesses of the MonoLT-Conf approach come from the fact that if the monolingual taggers are weak, their confidence scores are equally unreliable
- The results are switched between conditions LID-MonoLT (condition1) and MonoLT-LID (condition 2) for the two language pairs.

Discussion: Integrated conditions

- In general, except the ”COMB4:MonoLTSVM” condition all the INT conditions outperformed the COMB conditions.

For MSA – EGY:

- Adding more data helps, INT2:AllMonoData outperforms INT1:CSD, but combining the two conditions as training data, we note that INT3:AllMonoData+CSD outperforms the other INT conditions.
- **For SPA – ENG:**
- The worse INT condition is INT2:AllMonoData for Bangor (accuracy 78.57%) and INT1:CSD for Spanglish (accuracy 82.95%).

Discussion: Integrated conditions

For SPA – ENG:

- The largest gap in performance for Bangor could be due to a higher domain mismatch with the monolingual data used to train the tagger.
- Notable difference between the two language pairs is the significant jump in performance for the Bangor corpus from the **first three COMB** conditions from (68.35% to 96.31%).
- We observe a similar jump for the Spanglish corpus, the gap is much larger for the Bangor corpus

Discussion: Integrated conditions

- Some similar trends between the two combinations.
- MSA and EGY share a significant number of homographs some of which are cognates but many of which are not.
- The homograph overlap is quite limited in SPA-ENG.
- Adding the CSD to the monolingual corpora in the INT3:AllMonoData-CSD condition for MSA-EGY improves performance (1.5% absolute increase in accuracy).
- The results are not consistent across the SPA-ENG data sets.

Outline

- Introduction
  - ✓ Motivation.
  - ✓ Main Contribution.
- Approach
  - ✓ Monolingual POS Tagging systems
  - ✓ Combined Experimental Conditions.
  - ✓ Integrated Experimental Conditions.
- Evaluation
  - ✓ Datasets
  - ✓ POS Tag Sets
  - ✓ Results
- Discussion
  - ✓ Combined Conditions.
  - ✓ Integrated Conditions.
- Conclusion
  - Summary
  - Future Work

Conclusion: Summary

- Presenting detailed study of various strategies for POS tagging of CS data in two language pairs.
- The results indicate that depending on the language pair there are varying degrees of need for annotated code switched data in the training phase of the process.
- Languages that share a significant amount of homographs when code switched will benefit from more code switched data at training time. (e.g., MSA-EGY)
- Languages that are farther apart such as Spanish and English, when code switched, benefit more from having larger monolingual data mixed

Conclusion: Future Work

- All COMB conditions use either out of context or in context chunks as an input for the monolingual taggers.
- Our plan for the future work that process the out of context chunks to provide a meaningful context to the monolingual taggers.
- Extend the feature set used in the COMB4:MonoLT-SVM condition to include Brown Clustering, Word2Vec, and Deep learning based features

Outline

- Introduction
  - ✓ Motivation.
  - ✓ Main Contribution.
- Approach
  - ✓ Monolingual POS Tagging systems
  - ✓ Combined Experimental Conditions.
  - ✓ Integrated Experimental Conditions.
- Evaluation
  - ✓ Datasets
  - ✓ POS Tag Sets
  - ✓ Results
- Discussion
  - ✓ Combined Conditions.
  - ✓ Integrated Conditions.
- Conclusion
  - ✓ Summary
  - ✓ Future Work

Thanks !!