



MaChAmp: Multi-task Learning to the Rescue in Resource Scarce Scenarios



Benchmarks in Natural Language Processing (NLP)

THE WALL STREET JOURNAL.

© 1981 Dow Jones & Company, Inc. All Rights Reserved.

VOL. CXCVII NO. 14 *** Business Edition WEDNESDAY, JANUARY 21, 1981 *** FINANCIAL, NEW YORK 35 CENTS

After the Crisis

Torn U.S.-Iranian Ties Won't Heal for a While Despite Hostage Pact

Mutual Recriminations Likely, but in Long Run 'Normal' Relations Seen

Stay Apart, Let Time Pass

By KAREN BLUMER BROWN
Washington Post Staff Writer

Washington, D.C. (AP)—The long struggle between the United States and Iran to secure the release of 52 American hostages held in Tehran will have a long and bitter afterlife, but the two nations may find a way to coexist in a more normal relationship, says a senior State Department official.

The official, who spoke on condition of anonymity, said that while the United States and Iran will remain bitter enemies for some time, the two nations may find a way to coexist in a more normal relationship, says a senior State Department official.

What's News

Business and Finance World-Wide

THREE MAJOR BANKS reported mixed results for the fourth quarter. Citicorp, operating as Citicorp, reported a 10% increase in net income to \$1.1 billion, or 27 cents a share. Citicorp's profit rose 7% to \$1.1 billion.

First Pennsylvania Corp. had a fourth quarter loss of \$1.5 million, due mainly to a \$1.2 million expansion of loan loss provisions and \$1.5 million in other losses. The bank holding company's loss for the year totaled \$14.1 million.

American Express and the government may both see a victory in the "Flagler" dispute, resolved by the U.S. Supreme Court in favor of American Express. The court ruled that the government's seizure of the Flagler Hotel was unconstitutional.

Chesapeake Bell canceled a \$40 million note offering only a day before the securities were to be delivered to investors. The surprise move came at a time when the company's stock price had fallen 10%.

Jobs for Married Men

Year	Jobs (Estimated)
1977	100
1978	110
1979	120
1980	130

UNEMPLOYMENT among married men fell to 1.2% of the labor force in the fourth quarter, the Labor Department reports.

Live, From St. Paul, Here's 'A Prairie Home Companion'

A Lazy Town Humor Scores Big on Public Radio

Tax Report

A Special Summary and Forecast Of Federal and State Tax Developments

PERSONAL CONVENTIONS will not be held with later enactment. CONGRESS introduced legislation in 1978 by limiting one's deductible business expenses to 50% of the cost of the convention. A new law repeals the two-year limit, but only for conventions held after 1979.

Fire Hazard

Safety Officials Fear Skyscraper Holocaust Could Kill Thousands

They Cite Buildings' Design And Location, Lax Codes, Poison Gas From Plastic

Owners Note Record Is Good

Small text at the bottom of the page, including publication details and copyright information.

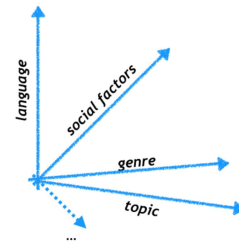
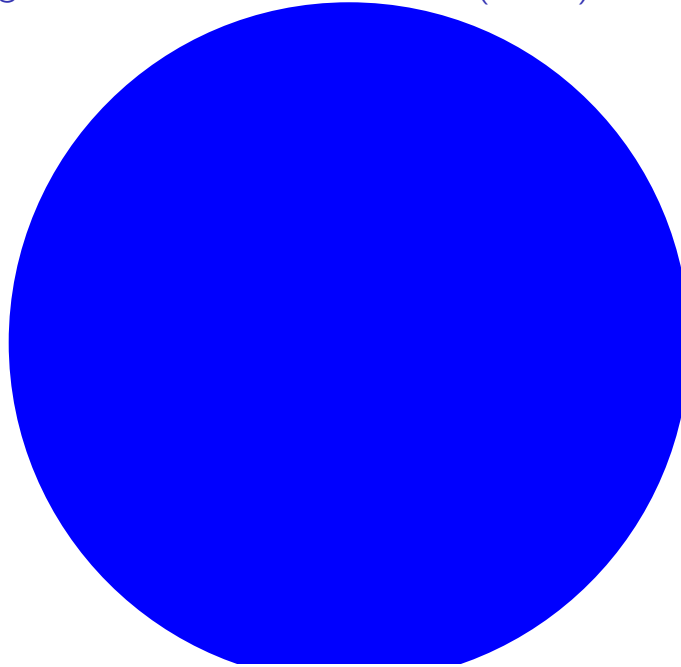
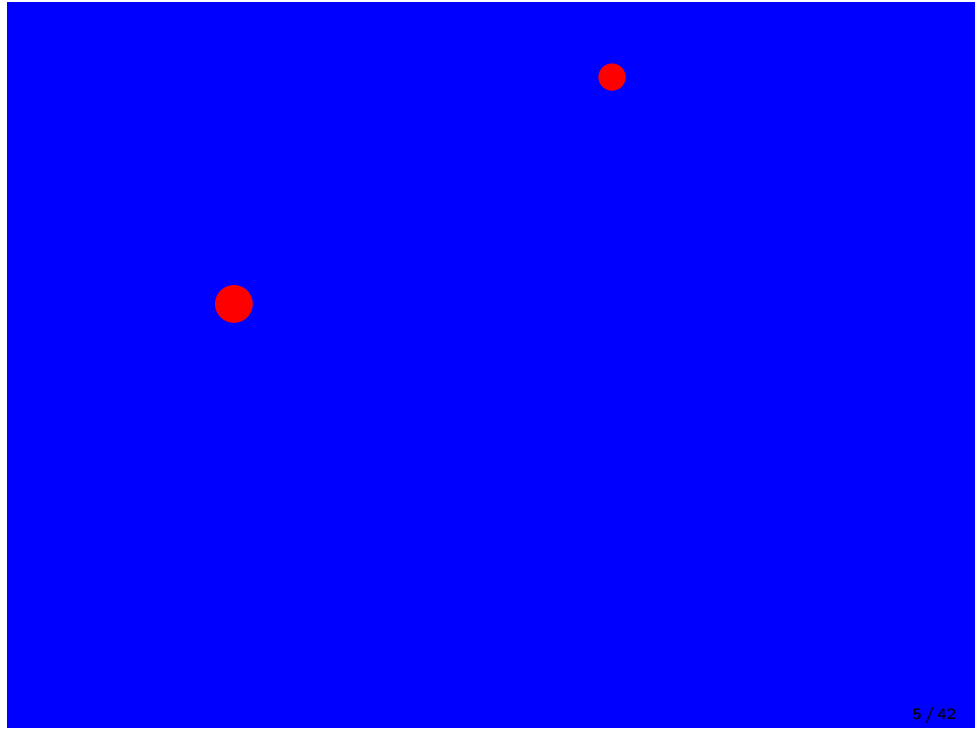


Figure 2: What's in a *domain*? Domain is an overloaded term. I propose to use the term *variety*. A dataset is a sample from the *variety space*, a unknown high-dimensional space, whose dimensions contain (fuzzy) aspects such as language (or dialect), topic or genre, and social factors (age, gender, personality, etc.), amongst others. A domain forms a region in this space, with some members more prototypical than others.

Language varieties that are annotated (in red)





What can we do?

- ▶ Annotate more?
- ▶ Cross-domain, cross-lingual learning



Multi-task learning to the rescue!

Standard in NLP:

- ▶ Pre-train a language model on raw data (billions of words)
- ▶ Fine-tune the language model on NLP-annotated data (thousands of words)

Framework: MaChaMp

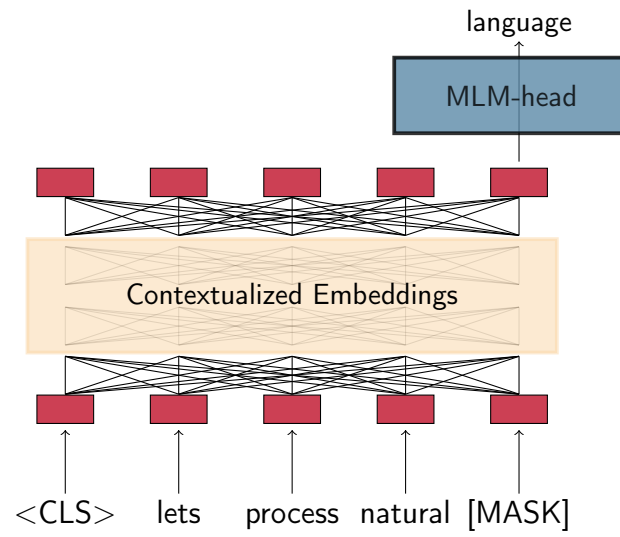
Massive Choice, Ample Tasks (MACHAMP):

 **A Toolkit for Multi-task Learning in NLP** 

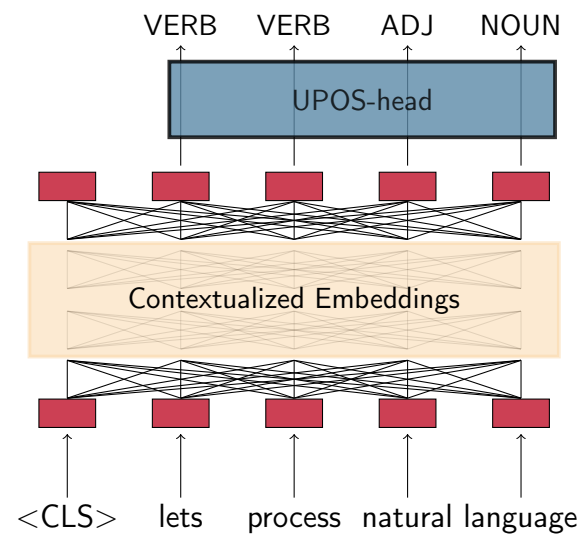
Rob van der Goot  **Ahmet Üstün**  **Alan Ramponi**   **Ibrahim Sharaf** 
Barbara Plank 

IT University of Copenhagen  University of Groningen  University of Trento 
Fondazione the Microsoft Research - University of Trento COSBI  Factmata 
robv@itu.dk, a.ustun@rug.nl, alan.ramponi@unitn.it
ibrahim.sharaf@factmata.com, bapl@itu.dk

MaChAmp




MaChAmp



- ▶ This is the default setup for all NLP tasks these days; sharing happens over time: MLM \Rightarrow TGT task
- ▶ MaChAmp can do much more!, we add multi-task learning after the first step

MaChAmp #001



Abilities:
 joint training loss weighing
 sequential training layer attention
 dataset embeddings dataset smoothing

MaChAmp is a multi-task NLP toolkit, it can seemingly effortlessly handle multiple NLP tasks simultaneously. It supports a wide variety of NLP tasks, and can easily handle multiple datasets at once.

More information on:
[machamp-nlp.github.io](https://github.com/machamp-nlp)

IT UNIVERSITY OF CPH

Examples of tasks	Input	Output
classification	Smell ya later!	negative
mlm	Gotta [MASK] em all	catch
multiclass	That will be 5\$	inform request
multisec	I never caught Snorlax	per:1 n:sin tens:past n:sin
regression	You're playing cats	1.2
seq	I want to be the best	PRN VB PART AUX DT ADJ
seq_bio	Ash from Pallet Town	Ash:PERS Pallet Town:LOC
tok	Gary, Gary, he's the man.	Gary , Gary , he 's the man .
dependency	Brock wants to fight	Brock wants to fight
		Rob van der Goot

Examples of tasks

Input	Output
classification	
Smell ya later!	negative
mlm	
Gotta [MASK] em all	catch
multiclas	
That will be 5\$	inform request
multiseq	
I never caught Snorlax	per:1 n:sin _ tens:past n:sin
regression	
You're playing cats	1.2
seq	
I want to be the best	PRN VB PART AUX DT ADJ
seq_bio	
Ash from Pallet Town	Ash:PERS Pallet Town:LOC
tok	
Gary, Gary, he's the man.	Gary , Gary , he 's the man .
dependency	
Brock wants to fight	Brock wants to fight



xSID: Cross-lingual Slot and Intent Detection

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi and Barbara Plank



Slot and Intent Detection

I'd like to see the showtimes for Silly Movie 2.0 at the movie house

Intent: SearchScreeningEvent

xSID

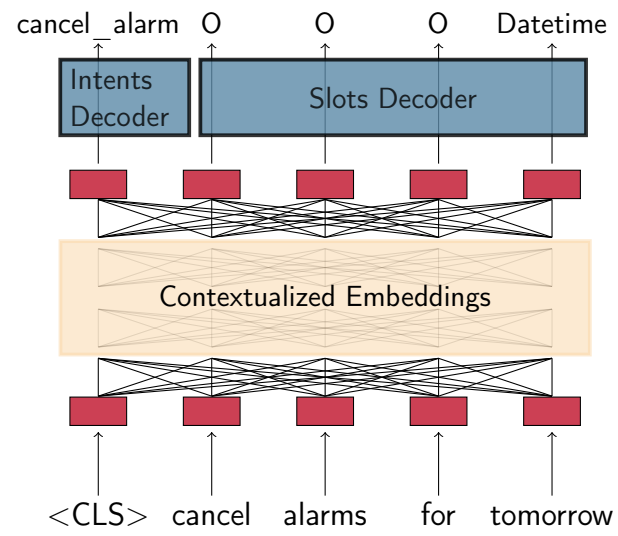
ar	أود أن أرى مواعيد عرض فيلم Silly Movie 2.0 في دار السينما
da	Jeg vil gerne se spilletiderne for Silly Movie 2.0 i biografen
de	Ich würde gerne den Vorstellungsbeginn für Silly Movie 2.0 im Kino sehen
de-st	I mecht es Programm fir Silly Movie 2.0 in Film Haus sechn
en	I'd like to see the showtimes for Silly Movie 2.0 at the movie house
id	Saya ingin melihat jam tayang untuk Silly Movie 2.0 di gedung bioskop
it	Mi piacerebbe vedere gli orari degli spettacoli per Silly Movie 2.0 al cinema
ja	映画館 の Silly Movie 2.0 の上映時間を見せて。
kk	Мен Silly Movie 2.0 бағдарламасының кинотеатрда көрсетілім уақытын көргім келеді
nl	Ik wil graag de speeltijden van Silly Movie 2.0 in het filmhuis zien
sr	Želela bih da vidim raspored prikazivanja za Silly Movie 2.0 u bioskopu
tr	Silly Movie 2.0 'in sinema salonundaki seanslarını görmek istiyorum
zh	我想看 Silly Movie 2.0 在 影院 的放映

Experiments

Baselines

- ▶ Baseline: contextualized embeddings with joint intent+slots

Baseline



Experiments

Baselines

- ▶ Baseline: contextualized embeddings with joint intent+slots
- ▶ Stronger baseline: translate training data to target language and map slot labels with attention (NMT-TRANSFER)

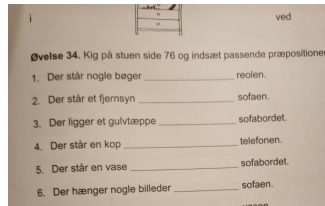
Experiments

Baselines

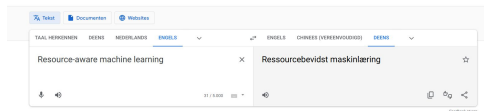
- ▶ Baseline: contextualized embeddings with joint intent+slots
- ▶ Stronger baseline: translate training data to target language and map slot labels with attention (NMT-TRANSFER)

New models:

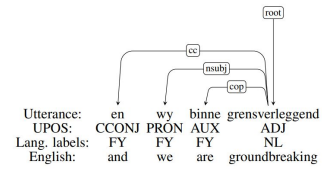
- ▶ Train on auxiliary task in target language:
 - ▶ Masked language modeling (AUX-MLM)
 - ▶ Neural machine translation (AUX-NMT)
 - ▶ UD-parsing (AUX-UD)



► MLM:



► NMT:



► UD-parsing:

Experiments

Evaluate 2 embeddings

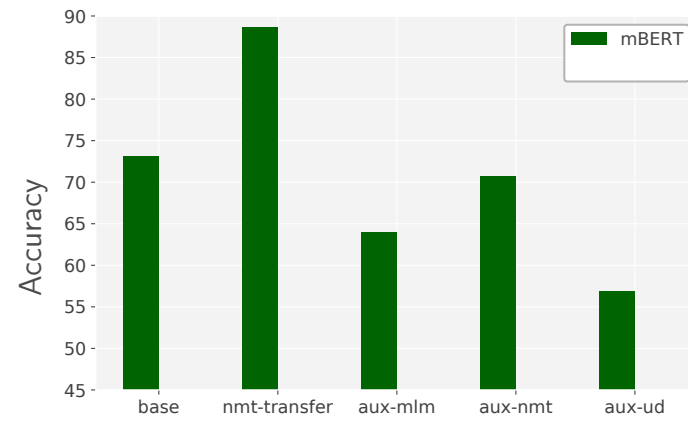
- ▶ mBERT: trained on 104 languages (12/13)
- ▶ XLM15: trained on 15 languages (5/13)

Results

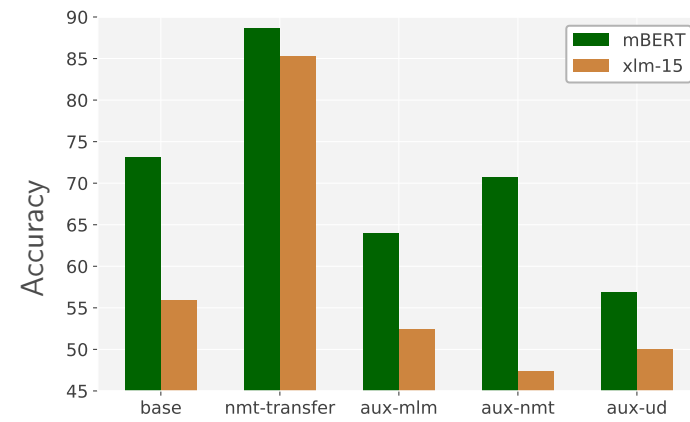
model	Time (minutes)
base	46
nmt-transfer	5,213
aux-mlm	193
aux-nmt	373
aux-ud	79

Table: Average minutes to train a model, averaged over all languages and both embeddings. For nmt-transfer we include the training of the NMT model.

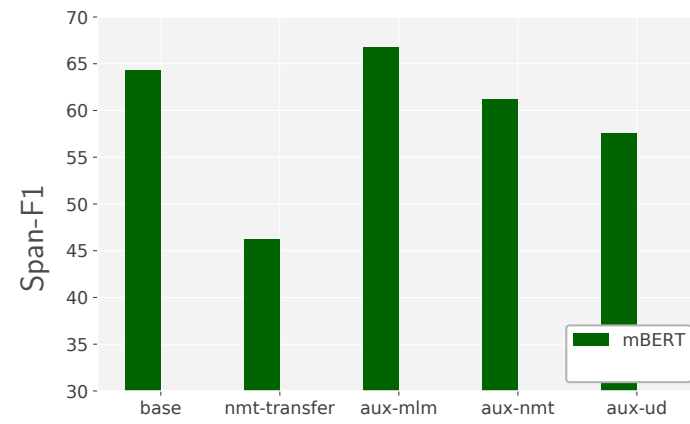
Results (intents)



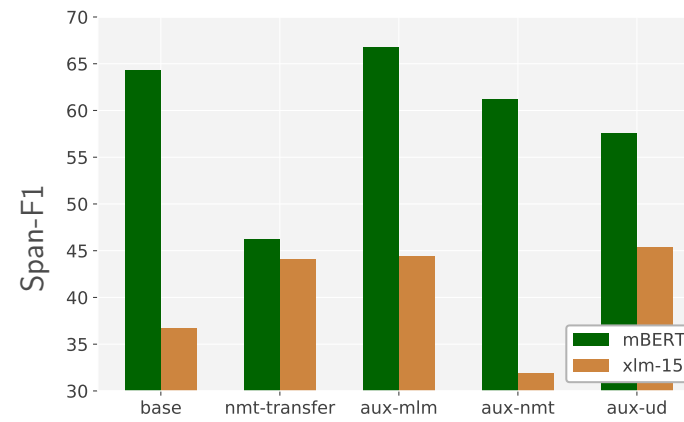
Results (intents)



Results (slots)



Results (slots)



Conclusions

Sentence level:

- ▶ NMT-transfer is hard to outperform, but costly
- ▶ Even baseline hard to beat

Span level:

- ▶ NMT-transfer performs bad (due to alignment)
- ▶ In-LM languages: only MLM helps
- ▶ Out-LM languages: More explicit tasks (UD) are faster and lead to better performance

Open questions

- ▶ Can NMT be used as auxiliary task?
- ▶ Are there better sentence level auxiliary tasks?
- ▶ Can NMT-transfer be improved with better word alignment?
- ▶ NMT and MLM hyperparameters
- ▶ Modeling jointly versus sequentially

How do we minimize memory in MaChAmp?

- ▶ It is based on language models, which are transformer-based.
- ▶ Transformer layers consider the whole input at once

Input to system is a batch of size 32*512:

- ▶ 32 sentences
- ▶ max 512 words: if more, we simply split up the sentence

We train a dependency parser on the English Web Treebank:

- ▶ 12,544 sentences; longest one 211 words
- ▶ Memory usage: 16GB!

We train a dependency parser on the English Web Treebank:

- ▶ 12,544 sentences; longest one 211 words
- ▶ Memory usage: 16GB!
- ▶ Goal: fit in 10GB

lets split up sentences after 128 words!:

▶ 16GB \Rightarrow 12GB!