



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 ★ ★ ★

ESTABLISHED 1889

WEDNESDAY, JANUARY 21, 1981

PRINCETON, NEW JERSEY

35 CENTS

After the Crisis

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

Mutual Recriminations Seen

Likely, but in Long Run

'Normal' Relations Seen

'Stay Apart, Let Time Pass'

By KIMBLE ELLIOTT BROWN

Staff Reporters of The Wall Street Journal
WASHINGTON—The long hostage ordeal that has so poisoned U.S.-Iranian relations is over, but its bitter residue will block American foreign-policy goals in Iran for the indefinite future.

After nearly 10 months of captivity, the 52 American hostages left Tehran yesterday for Algiers. In return, the U.S. released about \$1.1 billion in Iran's frozen assets in U.S. banks and their European branches by President Carter on Nov. 18, 1979.

That swap could eventually open the way for a more effective American policy to force the release of the hostages' relatives, see story on page 2.

check Soviet influence in Iran, laid a buffer between the Soviet Union and the Persian Gulf oil states. Meanwhile, President Reagan assumed office yesterday without the preoccupation of the hostage issue, which has shackled American foreign policy, especially in the Middle East, for over a year. Most of the world understood the nature of

What's News

Business and Finance

THREE MAJOR BANKS reported mixed results for the fourth quarter. Citicorp reporting net plunged 30% to \$96 million despite a \$37 million after-tax gain from the sale of real estate and lease residuals. But Manufacturers Hanover had a 10% increase to \$59.5 million, and Wells Fargo's profit rose 2% to \$34.1 million.

First Pennsylvania Corp. had a fourth quarter loss of \$68.8 million, due mostly to a \$12.5 million expansion of loan loss provisions and \$13.1 million to settle holder suits. The bank holding company's loss for the year totaled \$184.1 million.

Continued on page 2

American Telephone and the government may both see a victory in the "toughable corner" reached to settle their antitrust case. In return for divorcing itself of part of Western Electric and seven operating units, AT&T may be allowed to enter the data processing field.

Continued on page 2

Cincinnati Bell canceled a \$40 million note offering only a day before the securities were to be delivered to investors. The surprise move came after reports that a pending settlement of the company's antitrust suit

World-Wide

THE HOSTAGES FLEW to Algiers as President Reagan hailed their freedom.

Two Algerian jets left Tehran at 10:35 p.m. EST and refueled in Aden, carrying the former captives and Algerian diplomats. In Algiers, the 52 Americans were welcomed by U.S. legislators, and the group chafed in an airport limousine. A Soviet envoy said the hostages were leaving joyful Iran, where militants dominated the U.S.

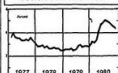
The release came after the Carter administration cleared up a late dispute over an appendix to the U.S.-Iranian accord. The U.S. agreed to waive \$1 billion of Iranian Iranian claims to the bank of England for transfer to Tehran. Washington's sanctions against Iran are to be lifted as well.

Antes said Carter spearheaded that Iran deflected the hostages' departure until after the change of Presidents in case his own men were emboldened.

Families cheered and wept at word of their captive relatives' return. Bells at the city hall in Albany, N.Y., rang 50 times at the military ceremony. In Alton, Ill., 44 sons, one second for each day the Americans were held prisoner.

CARTER PREPARED to greet freed captives at a hospital in West Germany. Having passed along the presidency, Carter briefly resumed home in Plains, Ga. "I doubt it at any time in our history more eagerly have I returned to my old home," Carter will be Reagan's emissary while three Americans recuperate at the U.S. mili-

Jobless Married Men



UNEMPLOYMENT among married men fell to 4.5% of the labor force in December from a revised 4.6% the preceding month, the Labor Department reports.

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

A Lazy Two Hours of Music And Small-Town Humor Scores Big on Public Radio

By LAWRENCE DINKOWSKI

and Reporter of The Wall Street Journal
LAKE WOBEGON, Minn.—If you've ever heard of this town, you know that it is the home of Ralph's Pasty Good Grocery, Our Lady of Perpetual Insignificance Church and the Shrine of the Unknown Norwegian.

And that it can't be found on any map but only on your radio for two hours every Saturday evening.

The mythical town of Lake Wobegon (pronounced wuh-bog-on) comes to life each week on "A Prairie Home Companion," a down-home radio show with a tiny Midwest-

Tax Report

A Special Summary and Forecast Of Federal and State Tax Developments

FOREIGN-CONVENTION rules are eased—with some exceptions.

Congress attached vacation allowances in 1976 by limiting one's deductible business conventions abroad to two a year and imposing spending and reporting requirements. A new law repeals the two-meeting limit, eases other restrictions and allows any foreign meeting that's "as reasonable" to have outside North America as within. However, since Richard Hanauer of Price, Waterhouse & Co., that means a meeting deductible under the old two-year rule may not be under the new reimbursement test.

That test involves the purposes and activities of the organization and the moving, the residences of the members, and the sites of their other gatherings. Meetings on cruise ships aren't deductible at all. The law broadens the deductible North American area to include Canada and Mexico, besides the U.S. and its possessions; that should stimulate meetings there.

The pending U.S.-Canada tax treaty would give the more lenient foreign-Minnesota rules Canada now has made trade decisions it never's been.

THEY'LL KEEP THE HOUSE and all the profits.

The price from selling your home normally is taxable, unless you put it into a new home by a certain time. But up to \$100,000 of gain on selling your principal residence will be excluded once from your gross income if you've ever lived and made your home there for five of the five years before 1978, 1979, 1980, and his wife, Joan, sold their home of 10 years in 1978.

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

By PHILIP R. HANSEN

Staff Reporters of The Wall Street Journal

Fire fighters say it more and over: to say one they can get to know: A major fire disaster is likely any day, a disaster that will rival an earthquake or flood in death and destruction.

Their nightmare is a fire that runs wild through a skyscraper during working hours, trapping thousands of people on upper floors. Hundreds, even thousands could be asphyxiated or burned to death in a few minutes.

Some think it's more than likely. "I'm not talking about probability. I'm talking about certainty," says Michael P. Lafferty, a former captain and 20-year veteran of the New York City Fire Department. He is currently safety officer at Citicorp's 59-story Manhattan office tower.

Why the Concern?

Speculators like Mr. Lafferty can tick off their concerns:

—More people than ever work in offices above their office buildings are going up, many in relatively flimsy structures.

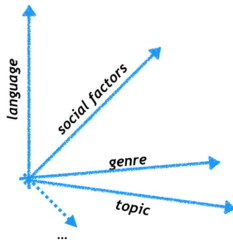
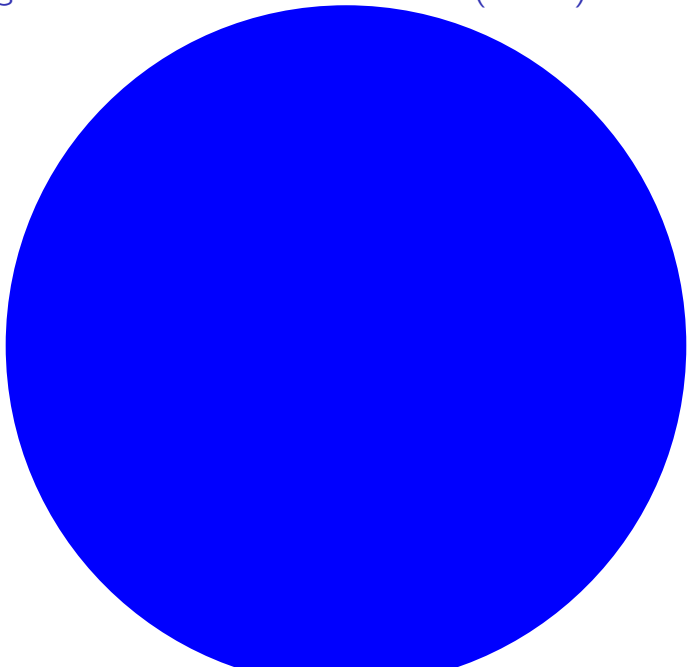


Figure 2: What's in a *domain*? Domain is an overloaded term. I propose to use the term *variety*. A dataset is a sample from the *variety space*, a unknown high-dimensional space, whose dimensions contain (fuzzy) aspects such as language (or dialect), topic or genre, and social factors (age, gender, 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 🇩🇪

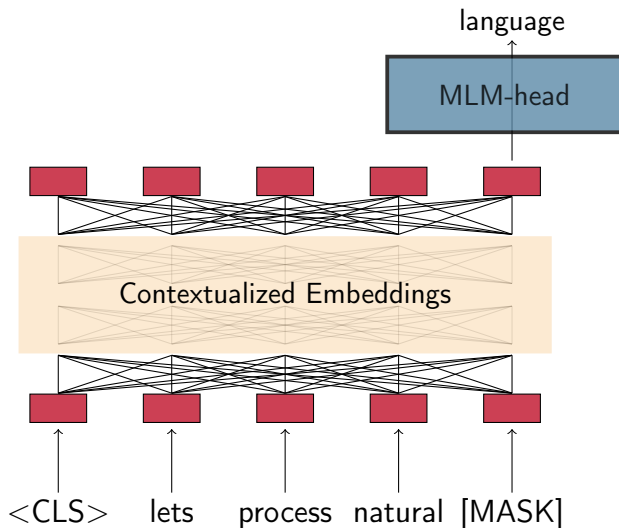
IT University of Copenhagen 🇩🇪 University of Groningen 🇳🇱 University of Trento 🇮🇹

Fondazione the Microsoft Research - University of Trento COSBI 🇮🇹 Factmata 🇪🇬

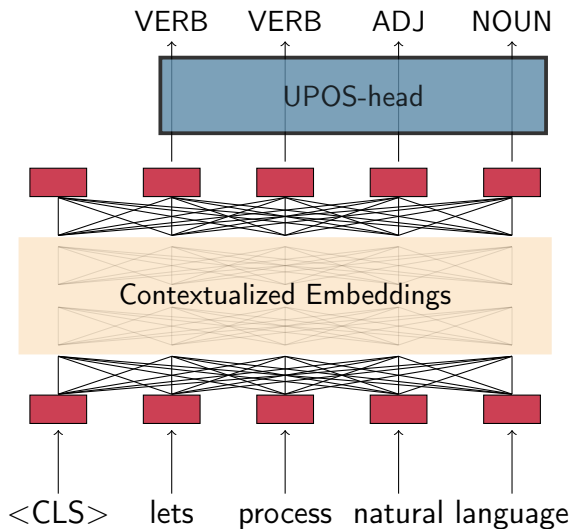
robv@itu.dk, a.ustun@rug.nl, alan.ramponi@unitn.it

ibrahim.sharaf@factmata.com, bapl@itu.dk

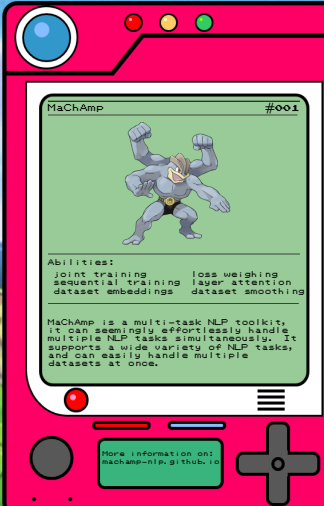
MaChAmp



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
- sequential training
- dataset embeddings
- loss weighing
- layer attention
- 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
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	

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

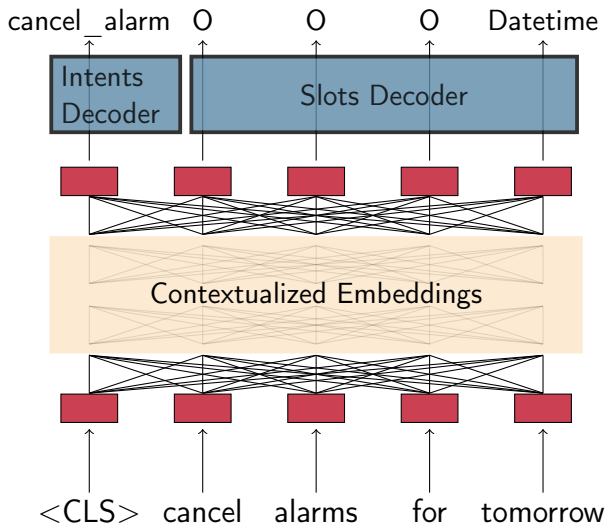
ar	أود أن أرى مواعيد عرض فيلم	Silly Movie 2.0	في دار السينما
da	Jeg vil gerne se spilletiderne for	Silly Movie 2.0	i biografen
de	Ich würde gerne den Vorstellungsbeginn für	Silly Movie 2.0	im Kino sehen
de-st	I mecht es Programm fir	Silly Movie 2.0	in Film Haus sechn
en	I'd like to see the showtimes 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'ın	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)

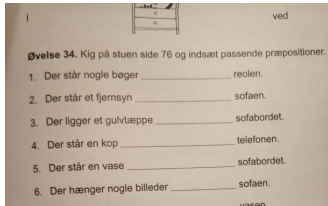
Experiments

Baselines

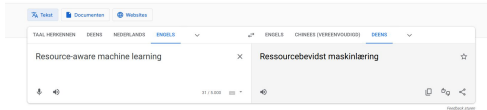
- ▶ Baseline: contextualized embeddings with joint intent+slots
- ▶ Stronger baseline: translate training data to target language and map slot labels with attention (NMT-TRANSFER)

New models:

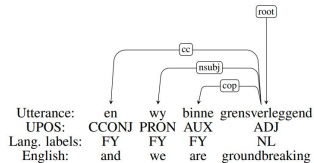
- ▶ Train on auxiliary task in target language:
 - ▶ Masked language modeling (AUX-MLM)
 - ▶ Neural machine translation (AUX-NMT)
 - ▶ UD-parsing (AUX-UD)



► MLM:



► NMT:



► UD-parsing:

Experiments

Evaluate 2 embeddings

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

Results

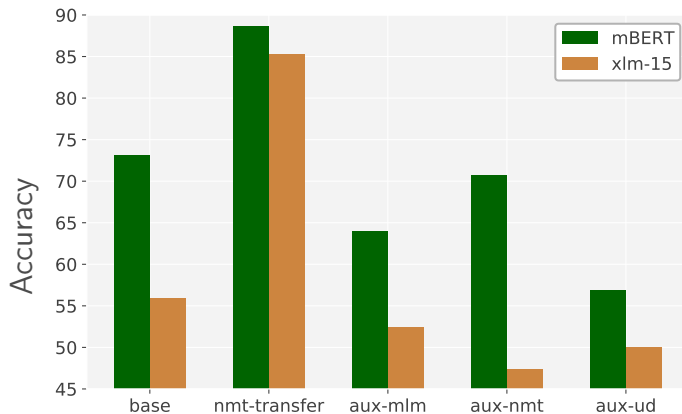
model	Time (minutes)
base	46
nmt-transfer	5,213
aux-mlm	193
aux-nmt	373
aux-ud	79

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)



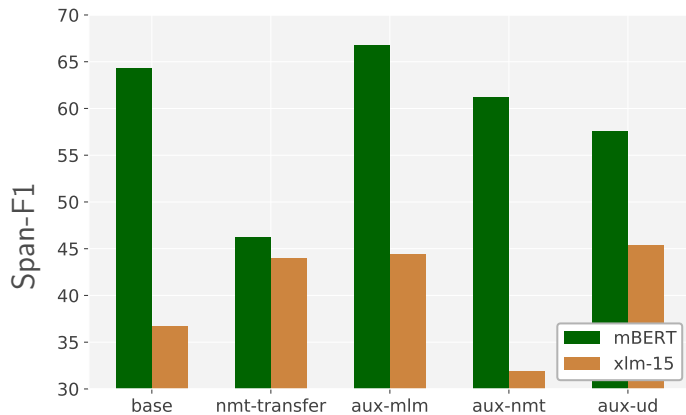
Results (intents)



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!

lets split up sentences after 128 words!:

- ▶ 16GB \Rightarrow 12GB!
- ▶ Note that splitting affects the input
- ▶ Effect on performance negligible
 - ▶ > 128 = not very frequent
 - ▶ Still has access to context in one direction (and context is longer for long sentences)

How to limit further?

- ▶ Batch size of 16 \Rightarrow lower performance
- ▶ Maximum words in batch (does this matter?)

Max 1024 words in batches of shape 32*128:

- ▶ 9.5GB!
- ▶ Probably because the batches are not all of size 32 anymore?

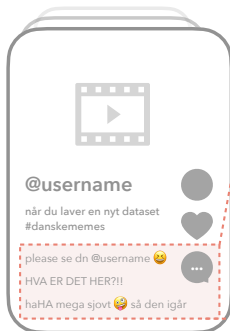
What about NLP for Danish?

- ▶ DaN+: Named entity recognition for Danish
- ▶ MultiLexNorm: lexical normalization of Arto data
- ▶ DanTok: POS tagging for Danish social media data

DanTok

- ▶ First NLP dataset on TikTok data
- ▶ First Danish social media POS tagging data
- ▶ Originally created for teaching: now in submission

DanTok:



Token	Norm	POS	Lang	Uncertain
please	Please	INTJ	CSType=INTRA Lang=en	0
se	se	VERB	Lang=da	0
dn	den	PRON	Lang=da	1
@username @username		PROPN	Lang=da	0
😊	😊	SYM	Lang=da	0

Research purpose:

- ▶ Investigate the effect of training data of POS tagger and MLM pre-training
- ▶ There is no Danish social media (only) trained LM or POS tagging training data

Model \ Data		+D-L TwB	-D+L DDT
-D+L	DANISH-BERT	49.60	77.98
	RØBÆRTA	60.82	70.17
	ÆLÆCTRA	63.30	74.20
+D-L	BERTWEET-B	38.00	67.90
	BERTWEET-L	36.55	67.40
	TWITTER-ROB	37.30	64.05
+D+L	TWITTER-XLM	77.15	72.15
	BERNICE	78.22	72.95
	TWHIN	81.38	72.65

Table: POS Tagging accuracy on the DanTok development set using combinations of in/out-of-domain (+D/-D) models and training data as well as in/out-of-language (+L/-L) models and training data.

Subset	Accuracy
All	85.9
Uncertain	62.1
Normalized	70.8
Unseen	83.3

Table: POS tagging accuracy for subsets of words

Conclusions:

- ▶ In-domain multilingual LM's with in-domain training data outperform Danish models!
- ▶ POS taggers struggle with the same cases as humans!

Tusind tak for i dag!