XFEVER:

Exploring Fact Verification across Languages

Yi-Chen Chang Canasai Kruengkrai Junichi Yamagishi

國立清華大學
NATIONAL TSING HUA UNIVERSITY

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

1. Introduction

1. Introduction

- Fact verification is part of the fact checking task, which is a process of verifying given claim is supported against a database of facts.

1. Introduction

- Fact verification is part of the fact checking task, which is a process of verifying given claim is supported against a database of facts.

Claim: Youtube is not a website.

Evidence: YouTube is an American video-sharing website headquartered in San Bruno, California.

1. Introduction

- Fact verification is part of the fact checking task, which is a process of verifying given claim is supported against a database of facts.

Claim: Youtube is not a website.

Evidence: YouTube is an American video-sharing website headquartered in San Bruno, California.

REFUTES

1. Introduction

- Problem Statement
  - Not realistic to collect data in all languages
    - Supervised learning models tend to be more accurate than unsupervised learning models
    - The existing data is usually in single language and most of them are English

1. Introduction

- Problem Statement
  - Not realistic to collect data in all languages
    - Supervised learning models tend to be more accurate than unsupervised learning models
    - The existing data is usually in single language and most of them are English
- Goal
  - Enable fact verification in other languages

1. Introduction

- Problem Statement
  - Not realistic to collect data in all languages
    - Supervised learning models tend to be more accurate than unsupervised learning models
    - The existing data is usually in single language and most of them are English
- Goal
  - Enable fact verification in other languages
- Hypothesis
  - Facts are facts regardless of language
  - i.e. The relationship between sentence pair in English and target language should be consistent

1. Introduction

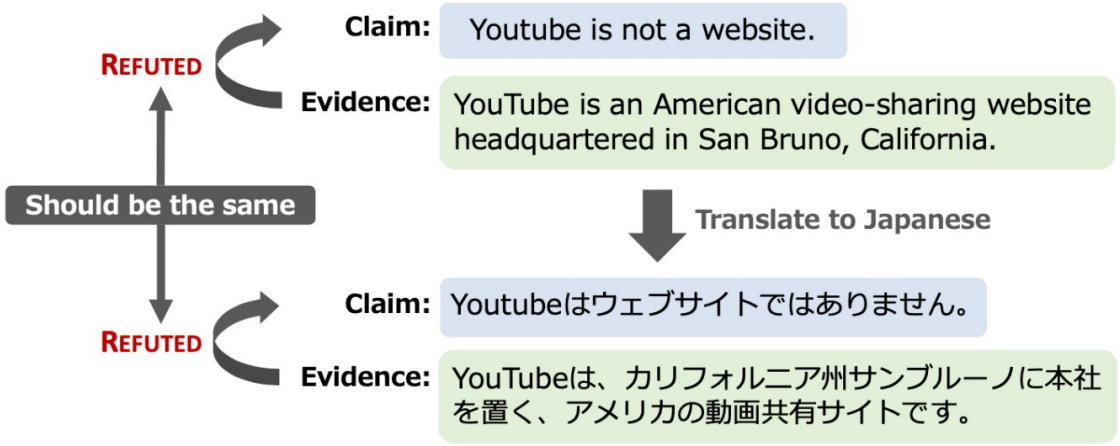


Figure 1: For the English example, it is clear that the given evidence refutes the claim. Suppose we have *accurate* translations from English to another language (e.g., Japanese). The claim in Japanese must also be refuted on the basis of the evidence in Japanese. In other words, the relationship between the claim and evidence text should be consistent across languages.

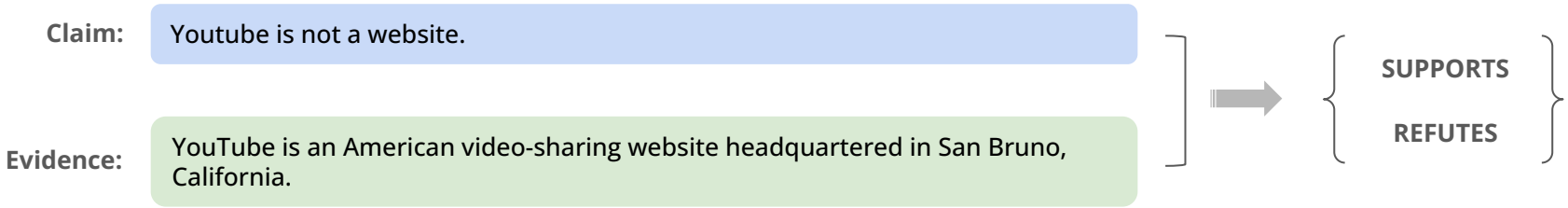
2. Related Work

2. Related Work: Automatic Fact Checking

- Recently, researchers have been exploring how fact checking can be automated to deal with the significantly increased information.

2. Related Work: Automatic Fact Checking

- Recently, researchers have been exploring how fact checking can be automated to deal with the significantly increased information.
- There are 3 stages:
  - Claim Detection
  - Evidence Retrieval
  - Claim Verification



2. Related Work: Fact-checking Databases

- FEVER [2]
  - A dataset for fact extraction and verification against textual sources.
  - It consists of 185K claims manually verified against the introductory sections of Wikipedia pages and classified as SUPPORTED, REFUTED or NOT ENOUGH INFO.
- WikiFactCheck [3]
  - A dataset of 124K examples extracted from English Wikipedia articles and citations
  - Consisting of claims and evidences which are both from the real world
  - There might be some limitation to apply to real-world via fact-checking system trained on manually generated data
- VitaminC [4]
  - A dataset with a total of over 400K claim-evidence pairs
  - The articles are collected from the most-viewed English Wikipedia pages as of January 2020 also including all articles in the FEVER dataset

3. Dataset

3. Dataset

- FEVER
  - A dataset for fact extraction and verification against textual sources
  - To focus on the fact verification task, we skip the extraction process by utilizing the processed dataset provided by Schuster et al. (2019) [5]
  - There are 3 classes: SUPPORTED (S), REFUTED (R) and NOT ENOUGH INFO (N)

3. Dataset

- XFEVER
  - Extend FEVER dataset to Cross-lingual FEVER dataset
  - Translate text from English into 5 target languages

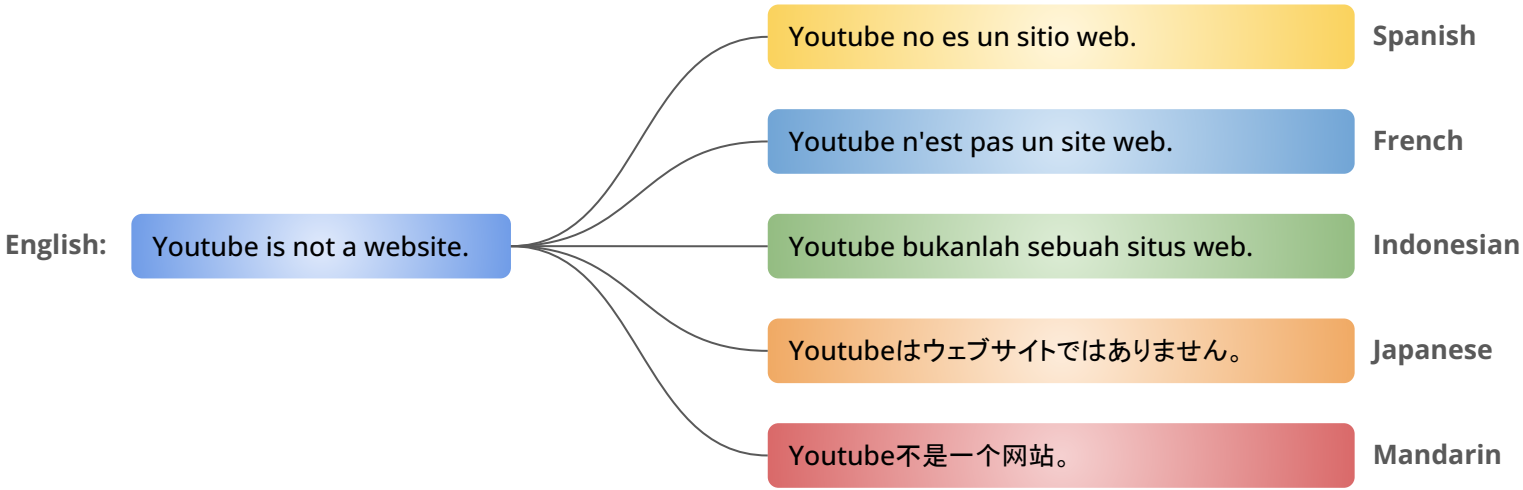


Figure 2: Extend FEVER dataset via translating data in English to other 5 target languages.

3. Dataset

Language	Claim / Evidence
English	Roman Atwood is a content creator. He is best known for his vlogs, where he posts updates about his life on a daily basis.
Spanish	
French	
Indonesian	
Japanese	
Chinese	

Table 1: Examples (claim and evidence) from six languages in the XFEVER dataset with the SUP class.

3. Dataset

Language	Claim / Evidence
English	Roman Atwood is a content creator. He is best known for his vlogs, where he posts updates about his life on a daily basis.
Spanish	Roman Atwood es un creador de contenidos. Es conocido sobre todo por sus vlogs, en los que publica a diario noticias sobre su vida.
French	Roman Atwood est un créateur de contenu. Il est surtout connu pour ses vlogs, où il publie quotidiennement des mises à jour sur sa vie.
Indonesian	Roman Atwood adalah pembuat konten. Dia terkenal karena vlog-nya , di mana dia memposting pembaruan tentang hidupnya setiap hari.
Japanese	ローマン・アトウッドは、コンテンツクリエイター。 彼は彼のブログで最もよく知られている、彼は毎日のように彼の人生についての更新を投稿している。
Chinese	罗曼-阿特伍德是一个内容创作者。 他最出名的是他的博客，在那里他每天都会发布关于他的生活的更新。

Table 1: Examples (claim and evidence) from six languages in the XFEVER dataset with the SUP class.

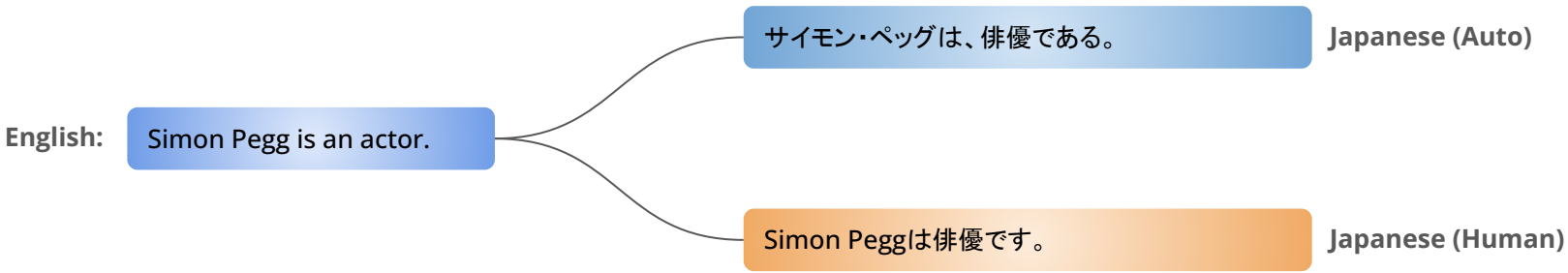
3. Dataset

Language	Claim / Evidence
English	Roman Atwood is a content creator. He is best known for his vlogs, where he posts updates about his life on a daily basis.
Spanish	Roman Atwood es un creador de contenidos. Es conocido sobre todo por sus vlogs, en los que publica a diario noticias sobre su vida.
French	Roman Atwood est un créateur de contenu. Il est surtout connu pour ses vlogs, où il publie quotidiennement des mises à jour sur sa vie.
Indonesian	Roman Atwood adalah pembuat konten. Dia terkenal karena vlog-nya , di mana dia memposting pembaruan tentang hidupnya setiap hari.
Japanese	ローマン・アトウッドは、コンテンツクリエイター。 彼は彼のブログで最もよく知られている、彼は毎日のように彼の人生についての更新を投稿している。
Chinese	罗曼-阿特伍德是一个内容创作者。 他最出名的是他的博客，在那里他每天都会发布关于他的生活的更新。

Table 1: Examples (claim and evidence) from six languages in the XFEVER dataset with the SUP class.

3. Dataset

- Translation Tools:
  - Auto Translation by DeepL
    - Apply to training, development, and testing dataset
  - Human Translation
    - Apply to testing dataset with randomly selected **600** examples only



4. Methods

- Scenario 1: Zero-shot Transfer Learning
  - Fine-tune on English dataset only
- Scenario 2: Translate-train Learning
  - Fine-tune on English as well as translated dataset

4. Methods: Scenario 1

- Zero-shot Transfer Learning
  - Fine-tune on English dataset only
  - Comparing the performance of language models pre-trained on monolingual and multilingual corpus(es).

Monolingual	# langs in corpus	Multilingual	# langs in corpus
BERT	1 (English)	mBERT	104
RoBERTa-base	1 (English)	XLM-R-base	100
RoBERTa-large	1 (English)	XLM-R-large	100

  - Training Object: $J_z(\theta) = \frac{1}{N} \sum_{(x,y)\in\mathcal{D}} L(x, y; \theta)$

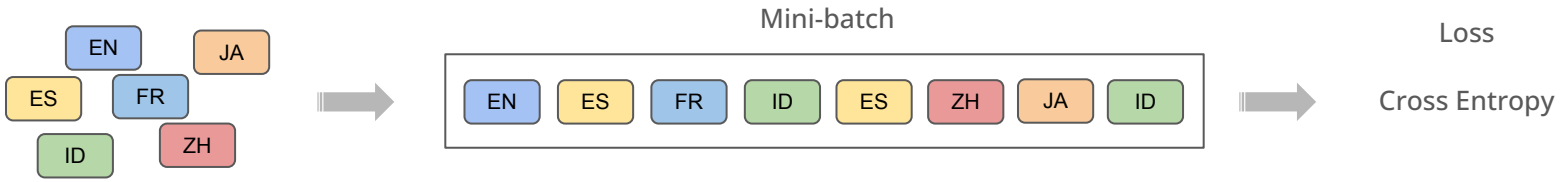
4. Methods: Scenario 2

- Translate-train Learning
  - Fine-tune on English as well as translated dataset
  - Different ways to fine-tune the pretrained language models
    - Parallel
    - Non-parallel

4. Methods: Scenario 2

- Translation-train Learning

  1. Non-parallel

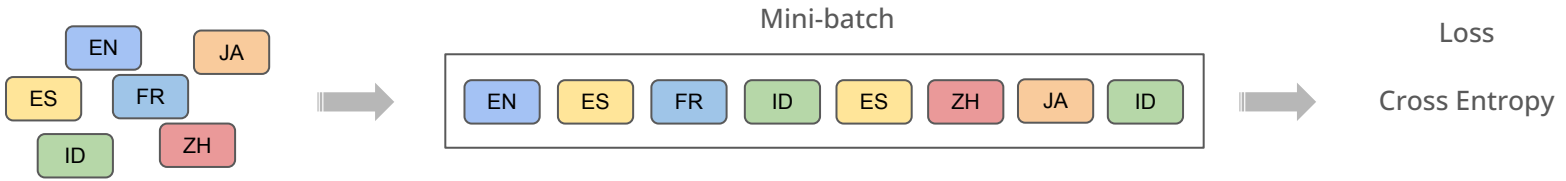


  2. Parallel

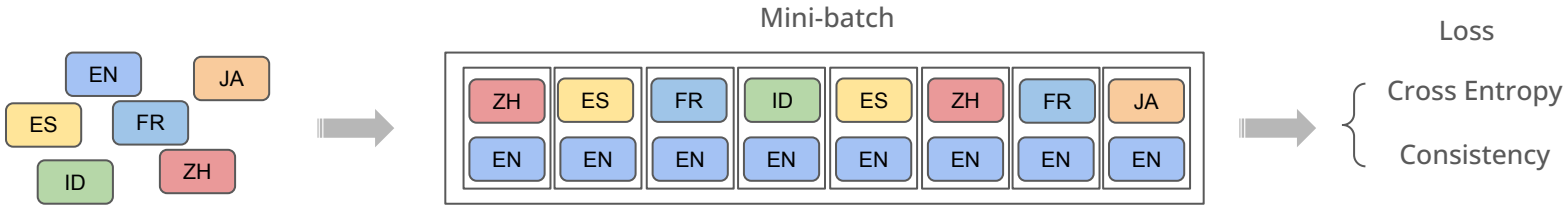
4. Methods: Scenario 2

- Translation-train Learning

1. Non-parallel



2. Parallel



4. Methods: Scenario 2

- Translation-train Learning

  1. Non-parallel

  Training Object: $$L = H(G,\ P_{lang}) = \sum_{x \in D \cup D'} G(x) \cdot \log \frac{1}{P_{lang}(x)}$$

  , where lang $\in$ {EN, ES, FR, ID, JA, ZH}

  2. Parallel

  Training Object: $$\begin{aligned} L &= L_{ce} + L_{consistency} \\ &= [L^S + L^T] + L_{consistency} \\ &= [H(G,\ P_{en}) + H(G,\ P_{lang})] + L_{consistency} \end{aligned}$$

4. Methods: Scenario 2

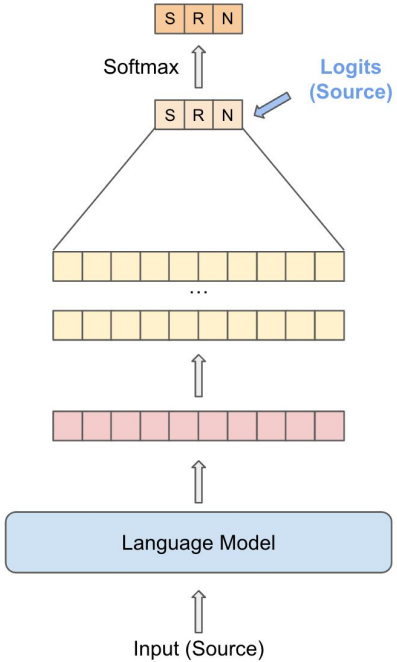
- Consistency Regularization
  - Prediction Consistency
  - Representation Consistency
- Consistency Loss
  - $L_{consistency} = \alpha_1 \cdot L_{pred} + \alpha_2 \cdot L_{re}$, where $\alpha_1,\ \alpha_2 \in \{0, 1\}$ and $\alpha_1 + \alpha_2 \in \{1, 2\}$.

4. Methods: Scenario 2

- Consistency Regularization
  - Prediction Consistency
    - Kullback–Leibler Divergence (KL)
      - A measure of how one probability distribution P is different from the other one
    - Symmetric KL (symKL)
      - The symmetric version of Kullback–Leibler divergence
    - Jensen–Shannon divergence (JSD)
      - The symmetrized and smoothed version of the Kullback–Leibler divergence

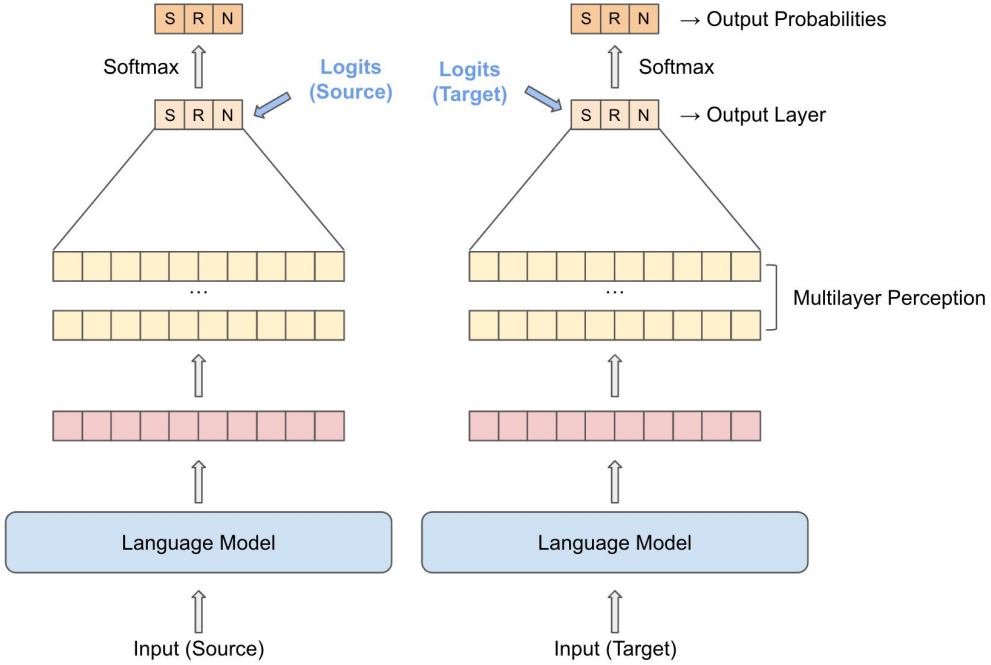
4. Methods: Scenario 2

- Consistency Regularization
  - Prediction Consistency



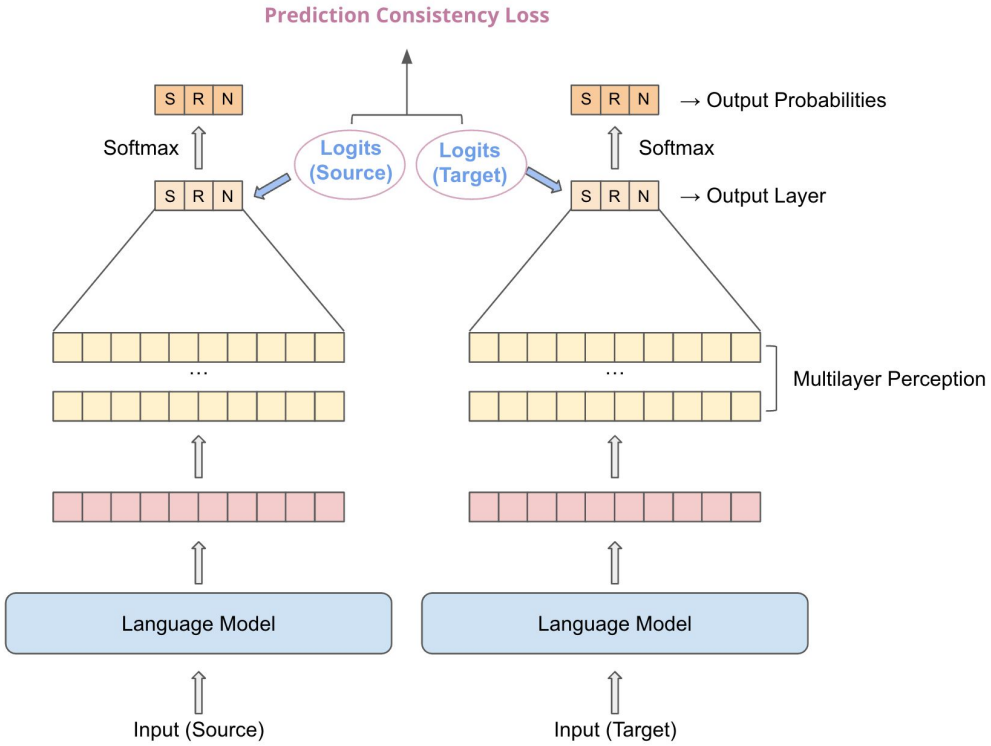
4. Methods: Scenario 2

- Consistency Regularization
  - Prediction Consistency



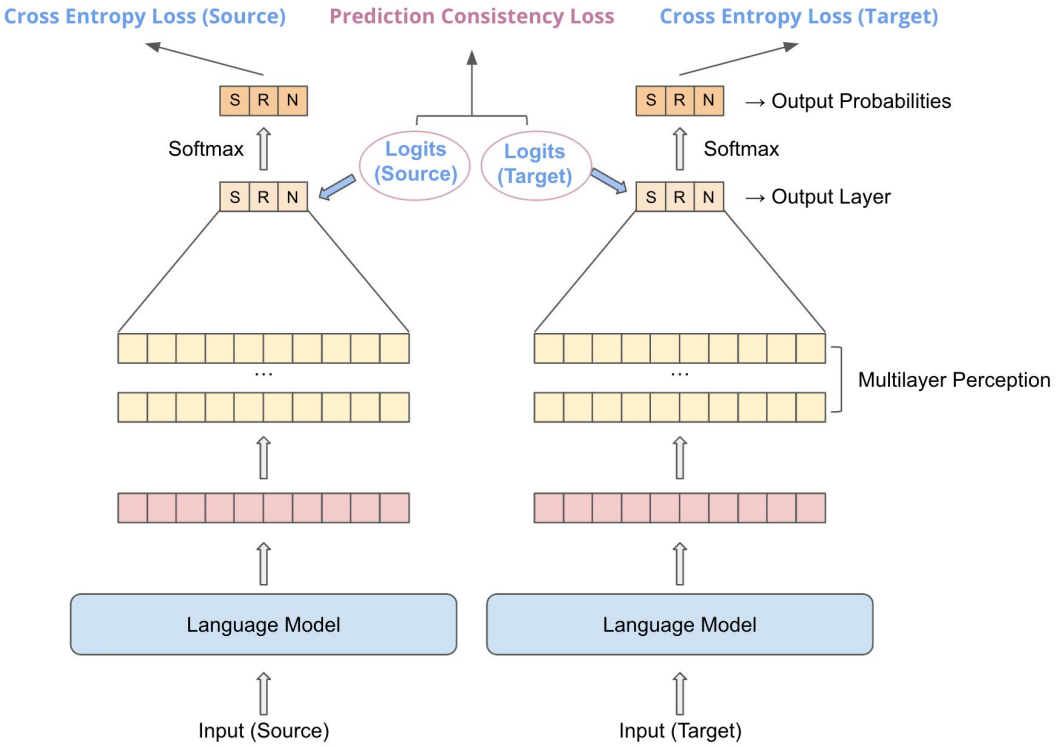
4. Methods: Scenario 2

- Consistency Regularization
  - Prediction Consistency



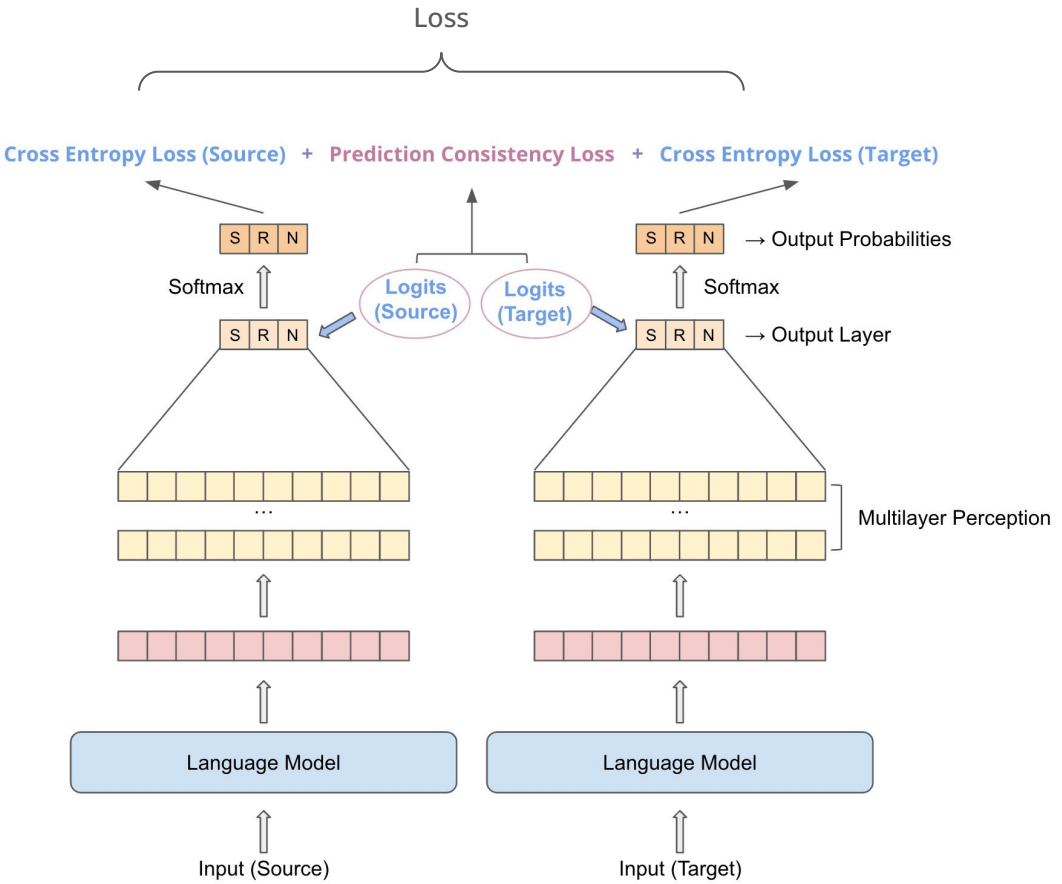
4. Methods: Scenario 2

- Consistency Regularization
  - Prediction Consistency



4. Methods: Scenario 2

- Consistency Regularization
  - Prediction Consistency



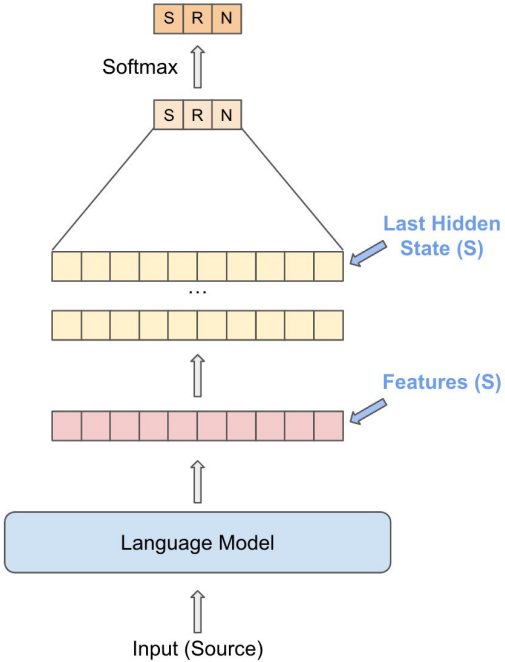
4. Methods: Scenario 2

- Consistency Regularization
  - Representation Consistency
    - Representations (r): Features / Last Hidden State
    - Functions
      - Mean Squared Error (MSE)
      - Negative Cosine Similarity

$$NegCosSim(r_{en}, r_{lang}) = 1 - CosSim(r_{en}, r_{lang})$$

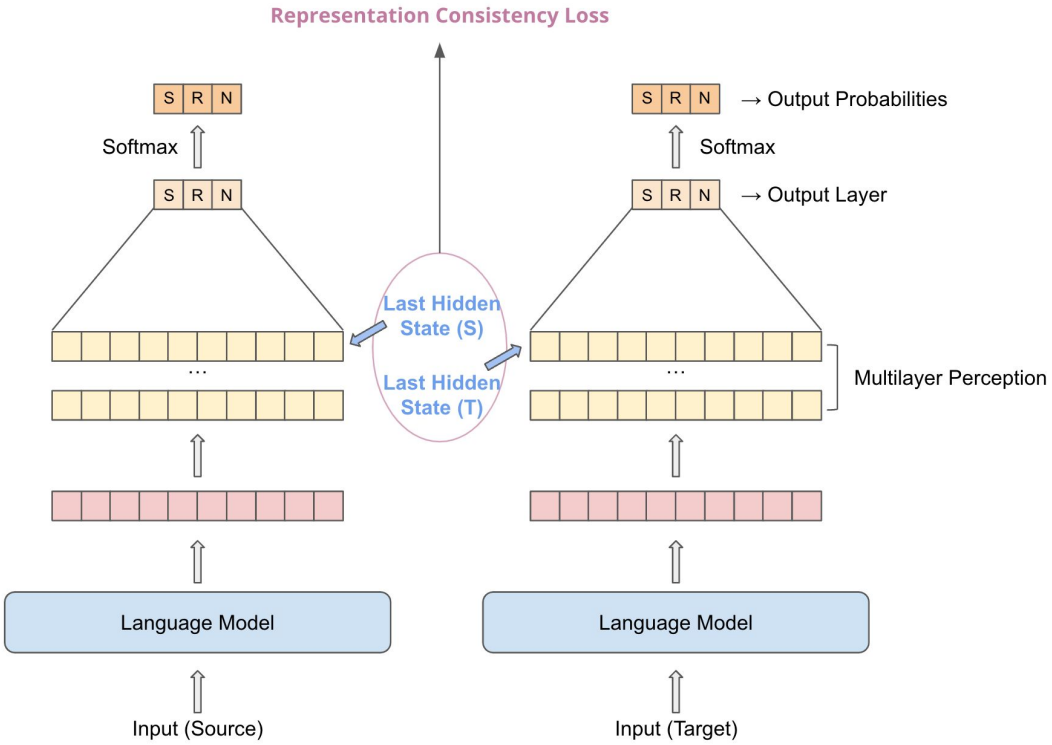
4. Methods: Scenario 2

- Consistency Regularization
  - Representation Consistency



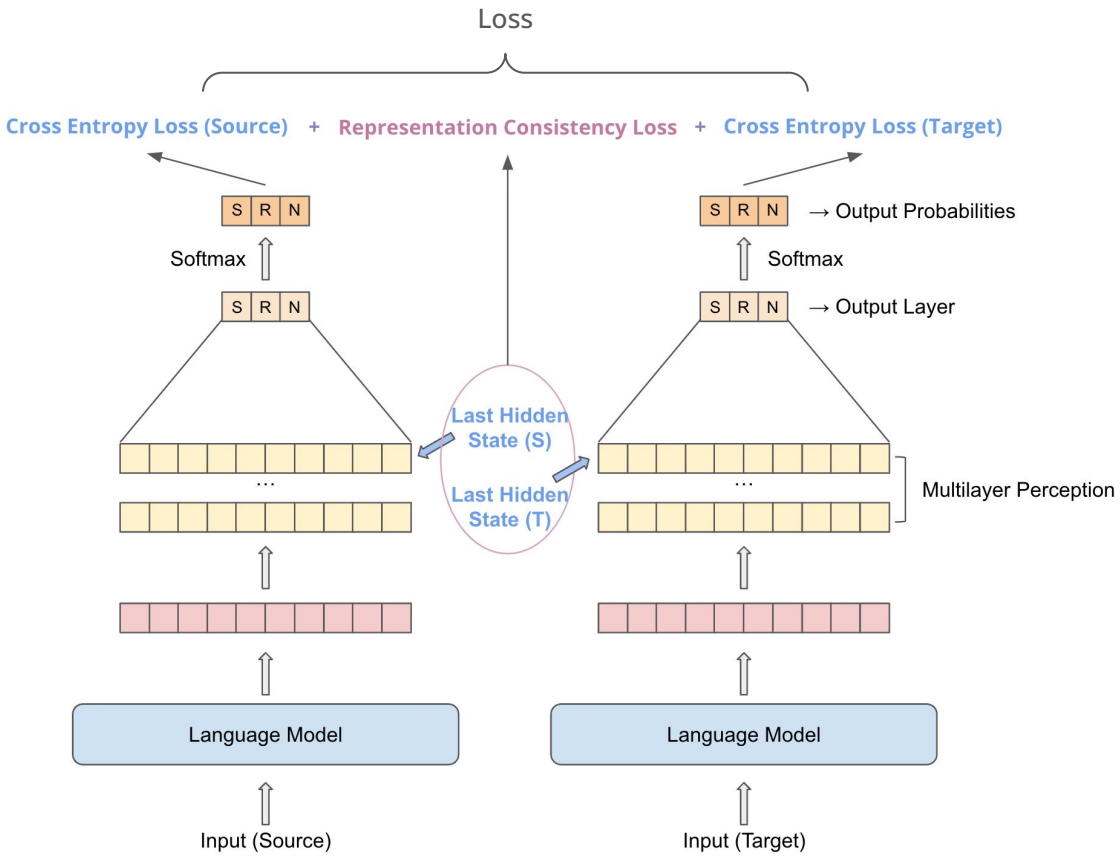
4. Methods: Scenario 2

- Consistency Regularization
  - Representation Consistency



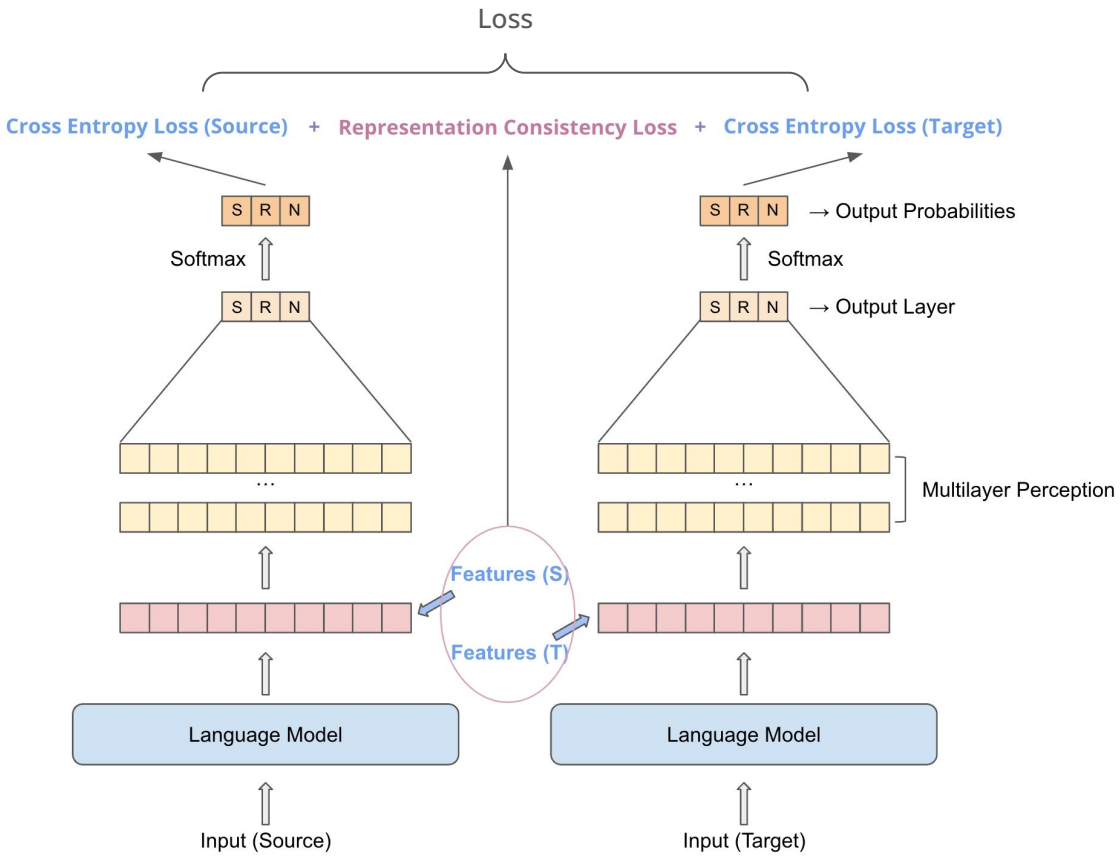
4. Methods: Scenario 2

- Consistency Regularization
  - Representation Consistency



4. Methods: Scenario 2

- Consistency Regularization
  - Representation Consistency



5. Experiments: Scenario 1

- Zero-shot Transfer Learning
  - Use language models pretrained on
    - Monolingual dataset
    - Multilingual dataset
  - Fine-tune on English dataset only
    - Both the training and development dataset are both in English only
  - Evaluate on English and other target languages

5. Experiments: Scenario 1

- Zero-shot Transfer Learning

Pretrained	en	es	fr	id	ja	zh	Accuracy Avg. (%)
Monolingual Pretrained Models							
BERT	87.26	56.24	57.46	54.56	36.51	38.55	55.10
RoBERTa-base	89.22	69.68	66.93	57.44	42.36	42.38	61.34
RoBERTa-large	90.65	77.63	71.65	56.52	41.32	42.28	63.34
Multilingual Pretrained Models							
mBERT	87.67	79.05	79.29	81.69	60.63	81.12	78.24
XLM-R-base	87.46	83.83	81.32	82.22	70.75	78.49	80.68
XLM-R-large	89.34	87.41	85.83	85.97	77.79	84.12	85.08

5. Experiments: Scenario 1

- Zero-shot Transfer Learning

Pretrained	en	es	fr	id	ja	zh	Accuracy Avg. (%)
Monolingual Pretrained Models							
BERT	87.26	56.24	57.46	54.56	36.51	38.55	55.10
RoBERTa-base	89.22	69.68	66.93	57.44	42.36	42.38	61.34
RoBERTa-large	90.65	77.63	71.65	56.52	41.32	42.28	63.34
Multilingual Pretrained Models							
mBERT	87.67	79.05	79.29	81.69	60.63	81.12	78.24 **(23.14 ↑)**
XLM-R-base	87.46	83.83	81.32	82.22	70.75	78.49	80.68 **(19.34 ↑)**
XLM-R-large	89.34	87.41	85.83	85.97	77.79	84.12	85.08 **(21.74 ↑)**

5. Experiments: Scenario 2

- Translation-train Learning
  - Use multilingual pretrained language models
  - Fine-tune on English and target language dataset
  - Evaluate on English and other target languages

5. Experiments: Scenario 2

(mBERT)		en	es	fr	id	ja	zh	Avg.
Non-Parallel								
-	-	88.20	86.40	86.29	86.43	74.18	86.28	84.63

5. Experiments: Scenario 2

(mBERT)		en	es	fr	id	ja	zh	Avg.
Non-Parallel								
-	-	88.20	86.40	86.29	86.43	74.18	86.28	84.63
Parallel								
KL	logits	87.21	85.76	85.35	85.60	82.13	84.40	85.07
symmetric KL	logits	86.87	85.68	85.20	85.58	81.78	84.32	84.91
JSD	logits	87.40	85.65	85.58	85.94	81.54	84.54	85.11
MSE	features	87.30	85.78	85.38	85.94	82.60	84.81	85.30
	penultimate layer	87.10	85.98	85.32	85.72	81.48	84.07	84.95
Negative Cosine Similarity	features	87.23	85.69	85.23	85.67	82.22	84.65	85.12
	penultimate layer	87.14	85.99	85.56	85.69	82.06	84.42	85.14

6. Discussions (1)

- Although the SOTA machine translation models can achieve quite good results, there might be some errors in the translated text.
- Evaluate the fine-tuned models on randomly selected 600 examples with 2 translation mechanisms.

6. Discussions (1)

- Better performances on auto translated text for scenario 2
  - It might be because the training data in target languages are obtained by auto translation
- Although the auto-translated texts might contain some errors, it doesn't affect the results of fact verification task too much.

(Scenario 1)	Source	Average	
mBERT	Auto	78.81	0.08
	Human	78.89	
XLM-R-large	Auto	84.64	0.36
	Human	84.28	

(Scenario 2)	Source	Average	
mBERT	Auto	84.39	0.28
	Human	84.11	
XLM-R-large	Auto	86.28	0.22
	Human	86.06	

6. Discussions (2)

- Ablation studies of different strategies for using the regularizations

Model		mBERT	XLM-R-large
Non-Parallel	-	84.63	**88.42**
Parallel	-	85.03	87.89
	w/pred. regularization	85.11	88.09
	w/re. regularization	**85.30**	87.81
	w/pred.+re. regularization	85.28	88.10

6. Discussions (3)

- Functionality of Prediction Consistency Regularization
  - Add confidence penalty to those wrong predictions with high probabilities
  - Metric: Expected Calibration Error (ECE)

	mBERT	XLM-R-large
Non-Parallel		
-	5.88	6.69
Parallel		
KL	4.69	6.89
symmetric KL	**2.00**	5.95
JSD	2.88	**2.95**

7. Conclusions

- Introduce a new benchmark XFEVER for cross-lingual fact verification task
- Evaluate and provide the baseline in 2 scenarios
    - Zero-shot transfer learning task
    - Translate-train learning task
- Study different consistency regularizations
    - Prediction Consistency Regularizations
    - Representation Consistency Regularizations
- Translation mechanisms don’t affect the results of cross-lingual fact verification task too much