|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nama, Tahun** | **Judul** | **Metode** | **Hasil** | **Gap** |
| Abdiansah, Azhari, & Sari, A.K. (2018) | INARTE: An Indonesian Dataset for Recognition Textual Entailment | Membuat dataset dengan *scraping* Wikipedia dan buku sejarah, dan dikonversi menjadi kalimat dengan *Automatic Hypothesis Generation* (AHG). Kemudian, kalimat hipotesis yang dihasilkan oleh AHG dipakai untuk menghasilkan kueri dengan *Query Generation* (QG) untuk dicari pada *search engine*.  Pada *source* yang dihasilkan oleh *search engine*, dilakukan *sentence tokenization & reduction* sehingga dapat menghasilkan pasangan kalimat teks-hipotesis.  Pelabelan *entailment* dilakukan dengan cara mengecek apabila hipotesis H merupakan *subset* dari teks T, maka akan bernilai *true*, selain itu bernilai *false*.  Akurasi dihitung dengan membagi jumlah pasangan *entailment* dengan jumlah pasangan teks-hipotesis. | *Dataset generation* menghasilkan 1.577 pasangan teks-hipotesis.  Apabila filter berdasarkan akurasi diterapkan, maka diperoleh jumlah pasangan *entailment* yang semakin sedikit. Hal ini dikarenakan oleh ketersediaan data pada web, kegagalan pada saat ekstraksi teks, kegagalan saat diproses oleh *Entailment Decision Algorithm* (EDA), dan perbedaan pada saat cek manual. | Perlu ada pengecekan apakah kalimat non-fakta dapat menghasilkan pasangan *entailment*. Terdapat kata-kata yang *overlapping* antara kalimat premis dan kalimat hipotesis. |
| Mahendra, R., Aji, A. F., Louvan, S., Rahman, F., & Vania (2021) | IndoNLI: A Natural Language Inference Dataset for Indonesian | Data *training* menggunakan *crowd annotator*. Teks premis dihasilkan dari Wikipedia, berita, dan artikel web. Untuk data yang bersumber dari berita, teks premis diambil dari *Indonesian* PUD dan GSD *treebanks* yang disediakan oleh Universal Dependencies 2.5 dan IndoSum. Untuk data yang bersumber dari artikel web, premis diekstrak dari situs web institusi dan blog.  IndoNLI menggunakan *single-sentence*, *double-sentence*, dan *multiple-sentences* pada premisnya.  Protokol anotasi terdiri dari 2 fase, yaitu penulisan hipotesis dan validasi label. Pada fase penulisan, tiap-tiap *annotator* diberikan 100-120 premis dan diminta untuk menuliskan 6 kalimat hipotesis untuk masing-masing premis (2 untuk setiap label: *entailment*, *neutral*, atau *contradiction*).  Untuk *expert annotator*, mereka diminta untuk menandai fenomena linguistik untuk melakukan inferensi pada masing-masing pasangan kalimat. *Expert annotator* memastikan pasangan kalimat hipotesis yang terbentuk memiliki fenomena linguistik yang terdistribusi.  Pada fase validasi, penulis melakukan verifikasi terhadap ~30% data yang dianotasi oleh *lay annotator* dan 100% data yang dianotasi oleh *expert annotator*. Validasi dilakukan dengan melabeli ulang setiap pasangan kalimat oleh 2 *independent annotator* lainnya. Jika label yang diberikan oleh kedua *independent annotator* sama dengan label awalnya, maka ditandai sebagai *gold label*.  Pemeriksaan *word overlap analysis* dilakukan dengan Jaccard *index* dan *Longest Common Subsequence* (LCS). Penulis juga melakukan penghitungan *new token rate*, yakni seberapa besar token hipotesis tidak muncul pada kalimat premis.  Selanjutnya dilakukan inferensi dengan model yang menggunakan Continous Bag-of-Words (CBOW) dengan *Indonesian fast-text embedding*. Model lain menggunakan Multilingual-BERT, IndoBERT, dan IndoBERT-lite.  Pengujian juga dilakukan dengan menggunakan kalimat hipotesis saja.  Untuk *task* antar bahasa, penulis menggunakan pendekatan *zero-shot*  *learning*, dan *translate-train*.  Pada *zero-shot learning*, model XLMR dilatih menggunakan potongan set data latih MNLI dan set data validasi XNLI.  Pada pendekatan *translate-train* data latih dan validasi MNLI diterjemahkan ke dalam bahasa Indonesia, kemudian model XLMR dilakukan *fine tuning* terhadap data yang telah diterjemahkan.  *Machine Translation* (MT) yang digunakan mengikuti arsitektur transformer standar, yaitu 6 layer *encoder* dan *decoder*. Model MT dilatih pada 13juta pasang korpus multi-domain. | *Dataset generation* menghasilkan 17.712 pasangan kalimat teranotasi. Semua data yang dianotasi oleh *expert* dan memiliki *gold label* digunakan menjadi sebuah set tes. Sementara data yang dianotasi oleh *lay annotator* dibagi menjadi set *development* dan tes. Data yang tidak termasuk pada fase validasi, dan pasangan data tanpa *gold label* yang dianotasi oleh *lay annotator* digunakan menjadi set *training*.  Data yang dianotasi oleh *lay annotator* memiliki tingkat *overlap* yang lebih tinggi dibandingkan dengan data yang dianotasi oleh *expert*. *New token rate* yang dihasilkan oleh *expert annotator* lebih tinggi, yang menunjukkan token yang digunakan lebih *diverse* ketika menghasilkan hipotesis.  Model dengan CBOW menghasilkan akurasi 50an% pada data *development* dan *test*. IndoBERT-lite menghasilkan 76,2% pada data *dev* dan 74,1% pada data *test lay*. IndoBERT-large menghasilkan akurasi 78,7% pada data *dev* dan 77,1% pada data *test lay*. Sementara mBERT menghasilkan akurasi yang sedikit lebih rendah dari IndoBERT. XLM-R memiliki akurasi lebih unggul, yaitu 85,7% pada set *dev* dan 82,3% pada set *test*. Hal ini terjadi dapat dikarenakan ukuran *pre-training* data bahasa Indonesia pada XLMR lebih besar daripada yang digunakan oleh IndoBERT-large.  Pengujian yang dilakukan dengan kalimat hipotesis saja memberikan hasil yang lebih jelek. Pada data *test lay* model terbaik (XLMR) menghasilkan akurasi ~60%, dan kinerja yang jauh lebih kecil pada data *test expert*, yakni 46%. Hal ini menunjukkan protokol untuk mengumpulkan data *test expert* mengurangi artifak anotasi secara efektif yang ada pada data *test lay*.  Untuk task antar bahasa, pendekatan *translate-train* memberikan hasil yang lebih baik pada set data *test expert* | Model masih terkendala menangani fenomena linguistik, terutama pada *numerical reasoning*, perbandingan, dan *superlative*. |
| Hu, H., Richardson, K., Xu, L., Li, L., Kuebler, S., Moss, L.S. (2020) | OCNLI: Original Chinese Natural Language Inference | Dataset dibuat tidak dengan menggunakan *Machine Translation* (MT), tapi dengan anotasi manual oleh *expert* dari *native speaker* bahasa Mandarin.  Protokol anotasi terdiri dari pemilihan premis, pembuatan hipotesis, dan verifikasi data.  Kalimat premis diambil dari dokumen pemerintah, berita, literatur, *talk show* TV, dan percakapan telepon. Genre yang dipakai juga memastikan keberadaan beberapa tingkat formalitas. Untuk *pre-processing*, simbol anotasi pada transkrip Callhome dihilangkan dan pemilihan premis dibatasi pada kalimat yang menggunakan 8 sampai 50 karakter.  Proses pembuatan hipotesis dilakukan dengan menggunakan *Multi-hypothesis elicitation*. Pada metode ini, *annotator* diminta untuk memberikan 3 kalimat untuk setiap label, sehingga menghasilkan total 9 kalimat.  Penulis lalu memberikan strategi MultiEncourage untuk membuat *annotator* memberikan hipotesis dengan memberitahu secara eksplisit tipe data apa yang dicari dan diberikan bonus tambahan apabila memenuhi kriteria.  Strategi lain dalam pembuatan hipotesis adalah MultiConstraint yang memberikan aturan bahwa hanya 1 dari 3 kalimat kontradiksi yang boleh mengandung kata negasi. Penulis lalu melakukan pengecekan secara acak, apabila ditemukan pelanggaran terhadap aturan, maka bayaran yang diberikan kepada *annotator* akan lebih kecil.  Untuk menangani masalah *word overlapping*, penulis memberikan aturan bahwa tidak boleh ada hipotesis yang *overlap* lebih dari 70% dengan premis.  Pada saat verifikasi data, pasangan kalimat dilabeli ulang oleh 4 *annotator* independen, kemudian dilakukan *majority vote* untuk memilih *golden label*.  Inferensi terhadap dataset yang telah dibangun dilakukan dengan menggunakan *Chinese-version* BERT-base dan RoBERTa-large. Implementasi menggunakan *benchmark* CLUE. | *Dataset generation* menghasilkan 6.000 pasang kalimat dengan pendekatan MultiEncourage dan MultiConstraint yang selanjutnya dibagi menjadi set data *dev* dan set data tes. Sementara sisa 50.000 data digunakan sebagai set data latih.  Inferensi terhadap dataset OCNLI dengan model tanpa transformer (seperti CBOW, dan biLSTM) memberikan hasil yang jauh lebih rendah dibandingkan dengan model yang menggunakan transformer. Model transformer terbaik, RoBERTa, menghasilkan akurasi sebesar 78,2%. Hasil ini ~12 poin dibawah akurasi kinerja manusia sebesar 90,3%.  Pada pengujian dengan data hipotesis saja, akurasi yang didapatkan adalah ~66%.  Kata dengan polaritas negative seperti “any” dan “ever”, kata negasi, dan “only” menjadi indikasi label *contradiction*. Kata “at least” memiliki bias terhadap label *entailment*. | Perlu ada teknik untuk mengurangi bias, dan *learning* dengan menggunakan *sentence representation* |
| Conneau, A., Rinott, R., Lample, G., Schwenk, H., Stoyanov, V., Williams, A., Bowman, S. R. (2018) | XNLI: Evaluating Cross-Lingual Sentence Representation | Pengumpulan data pada bagian bahasa Inggris dilakukan dengan mengumpulkan 250 kalimat dari masing-masing korpora *the Open American National Corpus: Face-to-face, Telephone, Government, 9/11, Letters, Oxford University Press* (OUP), *Slate*, dan *Verbatim*.  Penulis kemudian meminta *platform* *crowdsourcing* untuk menghasilkan 3 hipotesis untuk setiap premis, 1 untuk setiap label yang mungkin.  Proses pelabelan dilakukan dengan melabeli ulang oleh 4 *annotator* independen. Untuk setiap pasangan kalimat, sebuah *gold label* yang merepresentasikan hasil *vote* mayoritas kemudian diberikan.  Penulis kemudian melakukan penerjemahan dengan menggunakan jasa *One Hour Translation.* Kalimat-kalimat premis dan hipotesis diterjemahkan secara terpisah. Proses penerjemahan tidak mengganggu keterhubungan semantik antar bahasa.  Penulis selanjutnya melakukan pengembangan model terhadap dataset XNLI. Penulis mengevaluasi 2 tipe *cross-lingual sentence encoder*: (i) *pretrained universal multilingual sentence embedding* yang menggunakan rata-rata *word embedding* (X-CBOW), dan (ii) *Bidirectional-LSTM* (BiLSTM) *Sentence Encoder* yang dilatih pada data latih MultiNLI (X-BiLSTM). Kedua pendekatan menggunakan fungsi *alignment loss* yang sama.  Penulis melakukan ekstraksi vektor fitur dari BiLSTM, yaitu dengan cara menggunakan *hidden state* awal/ akhir, atau menggunakan nilai maksimum *overall* dari seluruh *hidden state*.  Pada pendekatan pertama, penulis menganggap *word embedding* fastText bahasa Inggris sebagai bagian *fixed*, dan melakukan *fine-tuning* *embedding* pada bahasa lain, sehingga rata-rata *word vector* pada kalimat dekat dengan *word vector* dalam bahasa Inggrisnya.  Pada pendekatan kedua, *sentence encoder* dilatih pada data latih bahasa Inggris MultiNLI bersama dengan *encoder* pada bahasa target, dengan harapan bahwa representasi dari kedua bahasa dekat di ruang *embedding*. Pada penelitian ini, menganggap *encoder* yang menghasilkan vektor dengan ukuran yang tetap sebagai *sentence representation*. |  | Perlu dilakukan *joint training* *encoders* dan *parameter sharing*. |
| Soudani, H., Mojab, M. H., Beigy, H. (2022) | Persian Natural Language Inference: A Meta-learning Approach | Penulis mengusulkan untuk menggunakan algoritma Meta-Learning. Meta-Learning merupakan proses pembangunan sebuah model yang dapat menyelesaikan sebuah *task* hanya dengan sedikit contoh data berlabel.  Penulis mempersiapkan pasangan kalimat dalam 2 bahasa untuk di*feed* ke meta-learner. Selanjutnya, penulis menggunakan 2 algoritma Meta-Learning berbasis optimisasi dan berbasis metrik, yaitu: *Reptile*, dan *Prototypical Network*.  Reptile merupakan algoritma ordo-pertama yang menggerakkan bobot menuju manifold dari rata-rata bobot pada parameter untuk *task* spesifik.  *Prototypical Network* merupakan algoritma berbasis metrik dari Meta-learning. *Prototypical networ*k mempelajari ruang metrik dimana klasifikasi dapat dilakukan dengan menghitung jarak ke representasi prototipe dari masing-masing kelas. Secara umum, algoritma ini terdiri dari *embedding network f* dan sebuah fungsi jarak *d(x1, x2)*. Fungsi jarak yang dipakai adalah *Euclidian distance*.  Penulis kemudian menggunakan DReCa sebagai *task augmentation strategy*. Tujuan dari DReCa adalah untuk mengambil sebuah *task* yang heterogen (sebagai contoh sebuah dataset) dan menghasilkan serangkaian *task* yang sudah terurai.  Penulis menggunakan dataset FarsTail sebagai dataset target dan XTREME sebagai dataset tambahan.  Penulis selanjutnya melakukan *fine-tuning pretrained model* menggunakan data latih dari bahasa Inggris dengan pendekatan *non-episodic*, lalu melakukan *fine-tuning* model yang didapat menggunakan data latih target. Pada pendekatan ini, penulis menggunakan mBERT, XLMR, dan ParsBERT.  Untuk pendekatan meta-learning, penulis menggunakan model XLM-R dengan layer *output* disesuaikan untuk tiap-tiap *task* dan dilatih dengan algoritma *Reptile* dan *Prototypical network*.  Selain itu, penulis juga melakukan evaluasi dengan metode *zero-shot*. | Model multi bahasa selalu lebih baik dibandingkan dengan model multi-task, dikarenakan NLI memerlukan representasi semantik yang lebih dalam.  Model dengan menggunakan Meta-learn menghasilkan kinerja yang lebih baik dibandingkan dengan *baseline* dan model *non-episodic*. Hal ini dikarenakan model meta-learning standar bertujuan untuk mencari model yang cukup umum untuk *fit* pada sebuah *task* yang baru.  Perbandingan algoritma Meta-learn yang digunakan, *prototypical network* memiliki kinerja lebih baik daripada *Reptile*. Hal ini dikarenakan *prototypical network* menggunakan representasi kelas alih-alih representasi contoh.  Apabila penulis menggabungkan data dari bahasa target dengan data dari bahasa tambahan (bahasa Inggris), maka akurasi model menurun. Hal ini dikarenakan data pada bahasa target berjumlah sangat sedikit dibandingkan bahasa lainnya.  *Episodic fine-tuning* secara signifikan menghasilkan akurasi yang lebih tinggi dibandingkan dengan *non-episodic fine-tuning*. Hal ini mendemonstrasikan bahwa *training* secara *episodic* efektif bahkan pada data bahasa tunggal dan dapat membuat *generality* pada level *training* dan *testing* data.  Penelitian dengan melihat *generality* dari parameter-parameter awal dilakukan dengan *Canonical Correlation Analysis* (CCA). Dari hasil penelitian, didapatkan model dengan Meta-learning menghasilkan CCA-*similiarity* yang lebih tinggi, yang mengindikasikan model mendapatkan parameter yang lebih *general* sebelum proses *fine-tuning*. |  |
| Kann, K., Ebrahimi, A., Mager, M., Oncevay, A., Ortega, J. E., Rios, A., Fan, A., Gutierrez-Vasquez, X., Chiruzzo, L., Gimenez-Lugo, G. A., Ramos, R., Ruiz, I. V. M., Mager, E., Chaudary, V., Neubig, G., Palmer, A., Coto-Solano, R., Vu, N. T. (2022) | AmericasNLI: Machine  translation and natural language  inference systems for  Indigenous languages of the  Americas | Penulis melakukan penerjemahan ke bahasa lokal dari bahasa Spanyol dengan menggunakan dataset XNLI. Bahasa lokal yang digunakan adalah Ashaninka, Aymara, Bribri, Guarani, Nahuatl, Otomi, Quechua, Raramuri, Shipibo-Konibo, dan Wixarika.  Beberapa kata yang tidak memiliki padanan dalam bahasa lokal, dibiarkan dalam bahasa Spanyol. Penulis kemudian memilih kalimat-kalimat yang relatif mudah untuk diterjemahkan. Dataset akhir berisikan terjemahan dari genre “*face-to-face*”, “*letters*”, dan *“telephone*” di XNLI. Kalimat-kalimat premis dan hipotesis diterjemahkan secara terpisah sehingga penerjemah tidak mengetahui padanan pada kalimat pasangannya.  Pada paper ini, penulis menyatakan bahwa *continued pretraining* merupakan bentuk adaptasi model yang paling dapat diandalkan untuk bahasa dengan sedikit sumber, karena tidak memperkenalkan *randomly initialized* parameter baru ke model.  Pendekatan lain adalah dengan berbasis terjemahan. Data paralel digunakan untuk melatih model *translation* dan untuk *fine-tuning*, dapat menerjemahkan data latih agar dapat *match* dengan bahasa yang akan dievaluasi (*translate-train*), **atau** dengan menerjemahkan data uji agar *match* dengan data latih (*translate-test*) | Hasil pengujian *zero-shot* dengan menggunakan XLM-R sangat rendah pada 10 bahasa lokal. Rata-rata akurasi ketika menggunakan bahasa Inggris sebagai bahasa saat *fine-tuning* adalah sebesar 38,48% dan 37,99% ketika menggunakan bahasa Spanyol.  Kinerja yang dihasilkan jauh diatas *baseline* dikarenakan banyak kata-kata bahasa Inggris atau Spanyol yang ditemukan pada data evaluasi.  Model adaptasi dengan *continued pretraining* menggunakan data tidak berlabel sangat berguna pada *task* ini. *Fine-tuning* pada bahasa Spanyol, dengan adaptasi, menambahkan rata-rata kinerja sebesar 5,86%, dan *fine-tuning* pada bahasa Inggris menambahkan rata-rata kinerja sebesar 5,22%. Ketersediaan data bahasa Quechua yang lebih banyak dibandingkan dengan bahasa lokal lainnya, berdampak pada kenaikan kinerja yang paling besar, yaitu sebesar 24,53% setelah menggunakan *continued pretraining*.  Penelitian ini juga memaparkan bahwa menerjemahkan data set latih mengungguli kinerja metode lainnya. Kenaikan kinerja antara *translate-train* dengan *zero-shot* *non-adapted* adalah sebesar 20,4% pada bahasa Quechua. |  |