মেশিন শেখার সাহিত্য পর্যালোচনা

(Machine Translation Literature Review)

দেবায়ন চক্রবর্তী, কজ স্যাম্বিও

(Debayan Chakraborty, Koj Sambyo*)

NIT, Arunachal Pradesh,Nharlagun,791112,INDIA
NIT, Arunachal Pradesh,Nharlagun,791112,INDIA

Abstract

এই কাগজ মেশিন অনুবাদ একটি সাধারণ ওভারভিউ প্রদান করে। এটি একটি সংক্ষিপ্ত ইতিহাস এবং মেশিন অনুবাদের মূল নীতিগুলি অনুসরণ করে এবং ভাষা মডেলগুলির একটি পরিসংখ্যান এবং পরিসংখ্যানগত মেশিন অনুবাদটির ধারণাগত ভূমিকা দ্বারা অনুসরণ করে।

KEYWORDS: FST, *SYNCHRONOUS CONTEXT FREE GRAMMAR,* WER, NIST,BLEU.

1. সূচনা

যন্ত্র অনুবাদ (এমটি) কম্পিউটারে ব্যবহৃত মেশিন অনুবাদ (এমটি) ব্যবহার করা হয় কম্পিউটারাইজড সিস্টেমে এক ভাষা, উৎস ভাষা (এসএল) থেকে পাঠ্য অনুবাদ, দ্বিতীয় ভাষাতে সমতুল্য পাঠ্য, লক্ষ্য ভাষা (টিএল)। এটি মানব অনুবাদক এবং সম্পাদকদের সহায়তায় বা ছাড়াও করা যেতে পারে, যদিও লক্ষ্যটি হ'ল ন্যূনতম মানবিক সহায়তা [৪]।

MT তে অনেকগুলি পন্থা রয়েছে যা বিভিন্ন ডিগ্রিগুলির ভাষা বিশ্লেষণের উপর নির্ভর করে। পরিসংখ্যানগত যন্ত্র অনুবাদ (এসএমটি) নতুন পদ্ধতির মধ্যে একটি এবং এমটি গবেষণার শেষ দশকের ফোকাস হয়েছে। এসএমটি সিস্টেমগুলি অন্যান্য এমটি সিস্টেমের বিপরীতে সমান্তরাল পাঠ্যের বৃহৎ সংস্থাগুলির পরিসংখ্যানগত বিশ্লেষণের উপর ভিত্তি করে তৈরি করে যা তাদের উন্নয়নের জন্য মানুষের জ্ঞানের উপর ব্যাপকভাবে নির্ভর করে [৪]।

2. ইতিহাস

17 শতকের পর থেকে পাঠ্যের স্বয়ংক্রিয় অনুবাদের ধারণাটি প্রায় [7]। অবশ্যই, এই মুহূর্তে গ্রন্থে পাঠ্যক্রমের জন্য প্রাথমিক শব্দ-অনুবাদ-শব্দ অনুবাদ ছিল এবং সম্পূর্ণরূপে কল্পিত ছিল।

আধুনিক কম্পিউটার সিস্টেমের উন্নয়নের পাশাপাশি ভাষাগত তত্ত্বের অগ্রগতিতে, স্বয়ংক্রিয় অনুবাদটি একটি বাস্তবতা হয়ে উঠেছে। এমটিতে গুরুতর কাজটি 1940 এর দশকের শেষ দিকে শুরু হয়েছিল এবং এটি গবেষণার একটি গুরুত্বপূর্ণ এবং ব্যবহারিক বিষয় হয়ে দাঁড়িয়েছে .1947 সাল থেকে 1960 এর দশকের শেষের দিকে এমটি সিস্টেমগুলিতে অনেক কাজ করা হয়েছিল। ভাষাতত্ত্বের অগ্রগতি গবেষকদেরকে এমন বিকাশের সুযোগ দেয় যা ভাষা বিশ্লেষণ এবং জড়িত বিশেষ ভাষার ভাষাগত বোঝার উপর নির্ভর করে।

1966 সালে অটোমেটিক ল্যাঙ্গুয়েজ প্রসেসিং অ্যাডভাইজরি কমিটি (এলএলএইচসি) এই সিদ্ধান্তে পৌঁছেছে যে MT অনুবাদকগণ [1] [1,7] হিসাবে দ্রুত বা সামর্থ্য হিসাবে গুণমানের অনুবাদ প্রকাশ করতে পারে না। তাদের প্রতিবেদনটি পরবর্তী দশকে মার্কিন যুক্তরাষ্ট্রের এমটি গবেষণায় পতন ঘটে। গবেষণা বিশ্বের অন্যান্য অংশে বিশেষ করে কানাডা, জাপান, এবং ইউরোপে অব্যাহত।
পরিসংখ্যানগত যন্ত্র অনুবাদের পরীক্ষায় 1 9 80 এর দশকের শেষের দিকে আইবিএমে [7] শুরু হয়। ইন্টারনেট এবং বৃহত্তর ভাষা ডেটাবেসগুলির কারণে সমান্তরাল পাঠ্যতার উপলব্ধি বৃদ্ধির কারণে 1990 এর দশকে এসএমটি সিস্টেমগুলির সম্ভাব্যতা বৃদ্ধি পেয়েছে। ২006 সালে মোস নামে একটি মুক্ত-উৎস এসএমটি টুলটি মুক্তি পেয়েছিল এবং বর্তমানে এটি সম্পূর্ণ সম্পূর্ণ এসএমটি সফ্টওয়্যার [3]।

3. মেশিন অনুবাদ

সাধারণ এমটি সিস্টেমে এসএল-তে একটি বাক্যের রূপান্তরকে টিএল-তে একটি অনুবাদগত সমতুল্য বাক্যের মধ্যে রূপান্তরিত করার নিয়মগুলির এক বা একাধিক সেট গঠিত হয়। [4]
আদর্শভাবে, একটি এমটি উভয় উৎস বাক্যের অর্থ সংরক্ষণ করা উচিত এবং লক্ষ্য বাক্য উৎপন্ন করার সময় টিএল এর ব্যাকরণগত এবং সিনট্যাকটিক নিয়ম অনুসরণ করা উচিত। এটি সাধারণভাবে কয়েকটি ধাপে সম্পন্ন হয় যেমন শব্দগুলির সরাসরি অনুবাদ, টিএল-তে কত শব্দ শব্দের একটি শব্দ বা বাক্যাংশের সাথে মিলিত হয় (এটি শব্দটির উর্বরতা হিসাবে পরিচিত) [4] এবং একটি দ্বিধান্বিত বা বিরক্তিকর শব্দ ডিআই প্রবৃত্তি অনুবাদ মধ্যে সিদ্ধান্ত।
এটি অর্জন করার জন্য, এমটি সিস্টেম বৃহত্তর দ্বিভাষিক (বা বহুভাষিক) অভিধানগুলি ব্যবহার করে যা ভাষাগুলির মধ্যে শব্দ অনুবাদের জন্য শব্দ সরবরাহ করে এবং নিয়মের বাক্য ও মডেলের একটি সেট এবং ক্রম অনুসারে শব্দগুলির উর্বরতা নির্ধারণ করে। [৪]। এই মডেলগুলি প্রাসঙ্গিক ভাষাগুলির স্বয়ংক্রিয় বা ম্যানুয়াল ভাষাগত বিশ্লেষণ দ্বারা নির্ধারিত হতে পারে।
এসএমটি সিস্টেম ইতিমধ্যে অনুবাদিত গ্রন্থে বা সমান্তরাল গ্রন্থে বড় সংস্থা পরিসংখ্যান বিশ্লেষণ ফলাফল অনুযায়ী শব্দগত অস্পষ্টতা সমাধান করে এই মডেল সাহায্য করার লক্ষ্যে। এটি সাধারণ SMT সিস্টেমের বিকাশের জন্য অনুমতি দেয়, যখন পর্যাপ্ত সমান্তরাল পাঠ্য সরবরাহ করা হয়, যে কোনও ভাষাগুলির মধ্যে অনুবাদ করতে অপেক্ষাকৃত দ্রুত প্রশিক্ষিত করা যেতে পারে [4]।
এমটি সিস্টেমে ব্যবহৃত সাধারণ মডেলগুলি নিম্নরূপ শ্রেণীবদ্ধ করা যেতে পারে I

3.1. *FST (FINITE STATE TRANSDUCERS)*

(FST) Finite State Automata (FSA) এর বৈচিত্র্য। একটি FSA রাজ্যের একটি সেট, লেবেল একটি সেট, এবং সংক্রমণ একটি সেট গঠিত। তারা রাজ্যের মধ্যে রূপান্তর হিসাবে FSAs লেবেল পড়া এবং লিখুন।
FSTs লেবেল দুটি সেট করে এই ধারণা প্রসারিত - একটি SL জন্য এবং এক টিএল জন্য। উৎস লেবেল থেকে একটি লেবেল পড়তে গেলে যথাযথ রূপান্তর করা হয় এবং টিএল সেট থেকে একটি লেবেল আউটপুট [4] তে লেখা হয়। ডিএসএফ এফএসটি মডেলের প্রকারের ধরন সেটের লেবেলগুলির ধরন অনুসারে শ্রেণীবদ্ধ।সাধারণত FST সিস্টেমগুলি একাধিক এফএসএস গঠিত হয় যা শব্দ বা বাক্যাংশ উভয় অনুবাদ এবং শব্দ বা বাক্যাংশগুলির পুনঃক্রমকরণের জন্য একত্রে যোগদান করে [4]।

3.1.1. ওয়ার্ড-ফর-ওয়ার্ডমডেলস:

ওয়ার্ড-ফর-ওয়ার্ড এফএসটি মডেলগুলি সবচেয়ে সহজ এমটি মডেল। অনুবাদটি শব্দ স্তরের উপর সম্পন্ন হয় এবং তারপর টিএল-এর শব্দটি টিএল-এর সিনট্যাক্স অনুসারে পুনঃক্রমিত হয়। এই মডেলের মধ্যে FST লেবেল শব্দ হয়। সাধারণভাবে প্রথম সিস্টেমে এফএসটি পৃথক শব্দ গ্রহণ করবে এবং তাদের উর্বরতা অনুসারে তাদের সদৃশ করবে।

পরবর্তী FST টিএল থেকে টিএল থেকে শব্দগুলি অনুবাদ করবে। সর্বশেষ FST শব্দগুলি পুনঃক্রম করবে [4]। ওয়ার্ড-ফর-ওয়ার্ড মডেলগুলি সাধারণত ব্যবহার করা হয় না কারণ তারা ফ্রেজ বা সিনট্যাক্স ভিত্তিক মডেলগুলির সাথে প্রতিদ্বন্দ্বিতা করতে পারে না ।

3.1.2. বাক্যাংশ-ভিত্তিক মডেলগুলি:

ফ্রেজ-ভিত্তিক মডেলগুলি শব্দটি সহজে শব্দ দ্বারা পরিবর্তিত বাক্যাংশগুলিতে শব্দভাণ্ডারে ভাগ করে শব্দ-জন্য-শব্দ মডেলগুলির উন্নতি করে। [3] এই এসএল বাক্যাংশগুলি অবশ্যই টিএল-তে একটি বাক্যাংশের সাথে মিলিত বা সংলগ্ন হওয়া আবশ্যক। ফ্রেজ অনুবাদ করা হয় এবং তারপর বাক্য একটি ফ্রেজ স্তর উপর পুনর্বিন্যস্তভাবে হয়। এই পদক্ষেপ প্রতিটি সিস্টেমের মধ্যে একটি পৃথক FST দ্বারা সঞ্চালিত হবে।

3.2. সিঙ্ক্রোনাস কনটেক্সট ফ্রি গ্রামার

একটি সমলয় প্রসঙ্গ-বিনামূল্যে ব্যাকরণ সিনট্যাক্স-ভিত্তিক মডেলের একটি প্রকার। সিনট্যাক্স ভিত্তিক মডেলগুলি বাক্যগুলিকে সংকীর্ণ প্রসঙ্গ-মুক্ত ব্যাকরণ (এসসিএফজি) হিসাবে সংজ্ঞায়িত মুক্ত ব্যাকরণগুলির একটি এক্সটেনশন (সিএফজি) হিসাবে ভাগ করে। একটি সিএফজি নন-টার্মিনালগুলির একটি সেট এবং টার্মিনালগুলির একটি সেট এবং প্রতিটি অ-টার্মিনাল থেকে টার্মিনাল এবং অ-টার্মিনালগুলির এক বা একাধিক ক্রম থেকে একটি ম্যাপিং সহ গঠিত। একক ব্যাকরণ পরিবর্তে, এসসিএফজিগুলির একই সময়ে দুটি ব্যাকরণ রয়েছে [4]। একটি এসসিএফজি-তে প্রতিটি অ-টার্মিনাল দুটি সিকোয়েন্সের সাথে ম্যাপ করা হয় - প্রতিটি ভাষার জন্য একটি। এই সিস্টেমটি একাধিক বাক্যাংশ, একক বাক্যাংশ এবং শব্দগুলির পুনর্বিন্যাসের জন্য অনুমতি দেয়। সিনট্যাক্টিক অংশগুলিতে অনুবাদ করা হয়েছে, যেমন বিশেষ্য বাক্যাংশ, ক্রিয়া বাক্যাংশ, পদান্বয়ী অব্যয়, ইত্যাদি। [৪]। এই সেগমেন্ট তারপর অনুবাদ এবং পুনর্বিন্যস্তভাবে করা যেতে পারে। এই পদ্ধতিটি স্বজ্ঞাত কারণ একটি ভাষার কাঠামো (এবং এইভাবে বাক্যটি ম্যানুয়ালি অনুবাদ করার সময় শব্দগুলির পুনর্বহাল) স্পীচ অংশগুলির পরিপ্রেক্ষিতে নির্ধারিত হয় এবং বিশেষ শব্দ বা সাধারণ বাক্যাংশ দ্বারা নয়।

4. পরিসংখ্যান ব্যবহার

Bayesian পরিসংখ্যান পদ্ধতি প্রায়ই উপরোক্ত মডেল পরিপূরক ব্যবহৃত হয় । কোন SMT একটি প্রদত্ত শব্দ বা বাক্যাংশের একটি সংখ্যাগুলির মধ্যে সিদ্ধান্ত নিতে হবে যখন সম্ভাব্যতা ব্যবহার করা যেতে পারে। Bayes থিওরেম নিচের সূত্র দ্বারা দেওয়া হয়:

$$P(A|B) = P(B|A)\cdot P(A)\ P(B)\ldots\ldots\ldots(1)$$

পরিসংখ্যানগত অনুবাদ সিস্টেমে A এবং B যথাক্রমে এসএল এবং টিএল-তে শব্দ, বাক্যাংশ, এমনকি বাক্যগুলিও থাকবে। P (B | A), P (A), এবং P (B) এর সম্ভাবনাগুলি সমান্তরাল পাঠ্যগুলিতে A এবং B এর ফ্রিকোয়েন্সি পরীক্ষা করে গণনা করা হয়। [4]

5. অনুবাদ মূল্যায়ন

বিভিন্ন এমটি সিস্টেমের মূল্যায়ন প্রাথমিকভাবে মানবিক মূল্যায়নকারীদের দ্বারা পরিচালিত হয়। তবে, সম্প্রতি বেশ কয়েকটি স্বয়ংক্রিয় সরঞ্জাম বিকশিত হয়েছে। অনুবাদগুলি সঠিকতার মূল্যায়ন করার জন্য এই সরঞ্জামগুলি ম্যাট্রিক্স বা পরিমাপ সরবরাহ করে [৪]। সাধারণভাবে একটি মেশিন-অনুবাদিত পাঠ্য সঠিকতাটি একই পাঠ্যকে তুলনা করে যা মানুষের দ্বারা অনুবাদ করা হয়েছে।

স্বয়ংক্রিয় মূল্যায়ন পদ্ধতি নিচে।

5.1. WER

ওয়ার্ড ত্রুটি হার (WER) কে স্বয়ংক্রিয়ভাবে একই উৎস বাক্যটির বিদ্যমান অনুবাদে রূপান্তরিত করার জন্য স্বয়ংক্রিয়ভাবে অনুবাদ করা পাঠ্যতে পরিবর্তনগুলির সংখ্যা গণনা করে গণনা করা হয়। এই পরিবর্তনগুলি প্রতিস্থাপন, সংযোজন এবং শব্দ মুছে ফেলার অন্তর্ভুক্ত [5]। WER মূল্যায়ন এর কিছু বৈচিত্র্য বাক্যটির বাক্যে অবস্থানের অবস্থানকে বিবেচনায় নেয় না, এই মূল্যায়ন পদ্ধতিগুলি অবস্থানকে স্বাধীন ত্রুটির হার বলা হয়।

5.2. NIST

এনআইএসটি মূল্যায়নের সিস্টেমটি সঠিকতা নির্ধারণ করতে অনুবাদ পাঠ্যে নেগ্রাম (এন এন শব্দগুলি যা ক্রম দীর্ঘ) ব্যবহার করে। স্বয়ংক্রিয়ভাবে অনুবাদিত পাঠ্যের এন-গ্র্যামগুলি একই উত্স বাক্যের ম্যানুয়ালি অনুবাদে তাদের সাথে তুলনা করা হয়। এন-গ্র্যামের সংখ্যার প্রচলনগুলি সাধারণ ভাষায় অনুবাদটির স্কোর নির্ধারণ করে [2]।

5.3. BLEU

একইভাবে NIST এর মূল্যায়নগুলিতে, BLEU একটি স্বয়ংক্রিয়ভাবে অনুবাদ করা বাক্যের জন্য একটি স্কোর গণনা করতে N-grams ব্যবহার করে। যাইহোক, BLEU ফাইনাল ন্যূনতম স্কোর গণনা করতে একটি জ্যামিতিক ফাংশন ব্যবহার করে [8]।

6. ভারতে আবেদন

ভারতের কথ্য ভাষার বিভিন্ন তালিকা রয়েছে। কমপক্ষে 30 টি ভিন্ন ভাষা এবং প্রায় 2000 টি উপভাষা চিহ্নিত করা হয়েছে। ভারতের সংবিধান জাতীয় সরকারের জন্য সরকারী যোগাযোগের দুটি ভাষা হতে হিন্দি ও ইংরেজী ব্যবহারকে নির্দিষ্ট করে দিয়েছে। উপরন্তু, এটি 22 টি নির্ধারিত ভাষাগুলির একটি সেট শ্রেণিবদ্ধ করে যা প্রশাসনিক ভাষাগুলির জন্য বিভিন্ন রাজ্যের দ্বারা আনুষ্ঠানিকভাবে গৃহীত হতে পারে এবং জাতীয় ও রাজ্য সরকারের মধ্যে যোগাযোগের মাধ্যম হিসাবে জাতীয় ভাষা পরিষেবা পরিচালনার জন্য অনুমোদিত হয়।

অনেক রাজ্যের নিজস্ব আঞ্চলিক ভাষা আছে। কেবল
প্রায় 5% জনসংখ্যা ইংরেজিতে কথা বলে। ভারতের মত একটি বৃহত্তর বহুভাষিক সমাজে, এক ভাষা থেকে অন্য ভাষায় নথির অনুবাদের জন্য একটি বড় চাহিদা রয়েছে। বেশিরভাগ রাজ্য সরকার নিজ নিজ আঞ্চলিক ভাষায় কাজ করে, অথচ কেন্দ্রীয় সরকারের সরকারী দলিলগুলি ইংরেজি বা হিন্দি হয়।

যথাযথ যোগাযোগের জন্য সংশ্লিষ্ট অঞ্চলের ভাষাগুলিতে এই দস্তাবেজ এবং প্রতিবেদনগুলি অনুবাদ করার প্রয়োজন আছে। আঞ্চলিক ভাষার সংবাদপত্রগুলি আন্তর্জাতিক সংবাদ সংস্থাগুলির কাছ থেকে প্রাপ্ত ইংরেজিতে সংবাদ অনুবাদ করতে হবে। একটি মেশিন সহায়তা অনুবাদ সিস্টেম মানব অনুবাদকদের দক্ষতা বৃদ্ধি হবে[6]।

References

1. ALPAC. Languages and machines: computers in translation and linguistics. a report by the automatic language processing advisory committee, division of behavioral sciences, national academy of sciences, 1966.
2. George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proceedings of the second international conference on Human Language Technology Research, HLT '02, pages 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.
3. Philipp Koehn. Statistical machine translation. http://www.statmt.org/.
4. Adam Lopez. Statistical machine translation. ACM Computing Surveys (CSUR), 40(3):8, 2008.
5. Franz Josef Och. Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, pages 160–167. Association for Computational Linguistics, 2003.
6. Sudip Naskar, Sivaji Bandyopadhyay. Use of Machine Translation in India: Current Status

7. TAUS. A translation automation timeline. http://www.translationautomation.com/timeline/a-translationautomation-timeline, 2013. [Online; accessed 27-April-2013].
8. Aurelia Drummer (DRMAUR002) Department of Computer Science, University of Cape Town. Literature Review: Machine Translation