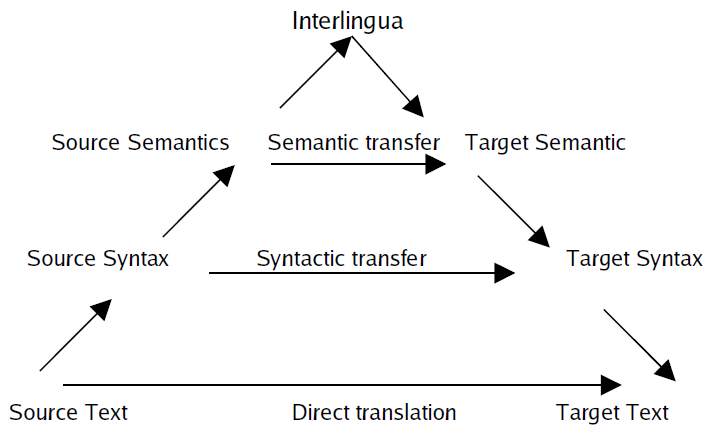
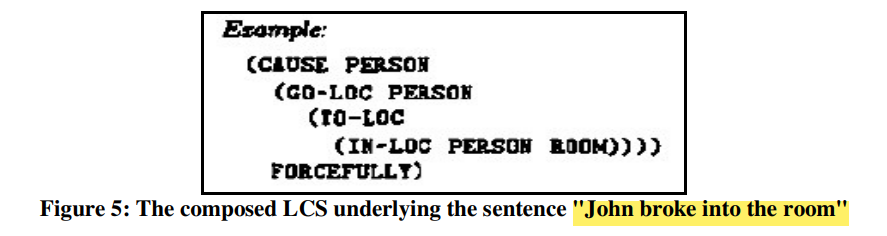
通常来说，机器翻译领域公认的经典（rule-based）方法有三种，direct，transfer和interlingua。前两者在课上都有介绍，这里不再赘述。第三种“基于中间语言的机器翻译”建立在两个理论（或者说信念）基础上：第一，机器翻译必须超越纯粹的语言信息（语法和语义），并涉及对文本内容的“理解”。早在上世纪六十年代机器翻译刚刚兴起时，Halliday提出自然语言中code和message是一体的，如果不能把meaning从text中剥离并转换成无歧义的机器表示，真正触及语言本质的翻译便难以实现。第二，长期以来，语言学家达成了一定的共识，即虽然各种语言“表面结构”差别很大，但它们有一个共同的“深层结构”。基于这一理念，科学家希望构建一种能够传达这种深层结构的普遍表征，这种表征具有自然语言所缺乏的规律性和易于预测。中间语言（interlingua）就是这样一种表征。

假设（理想状态下）确实对世界上所有的自然语言存在这样一种通用的、无歧义的中间语言表征，那么它在机器翻译中的应用很容易想象：其他基于规则的翻译通常需要在每对语言之间编写规则，而基于中间语言的翻译分为两个单语部分：将源语言文本分析为抽象的、普遍的、独立的中间语言表示，以及根据中间语言表示按词汇单位和句法结构生成目标语言文本。这样一来，只需要对N种语言分别编写N套分析和生成规则，就可以实现在任意两种语言之间的互译。

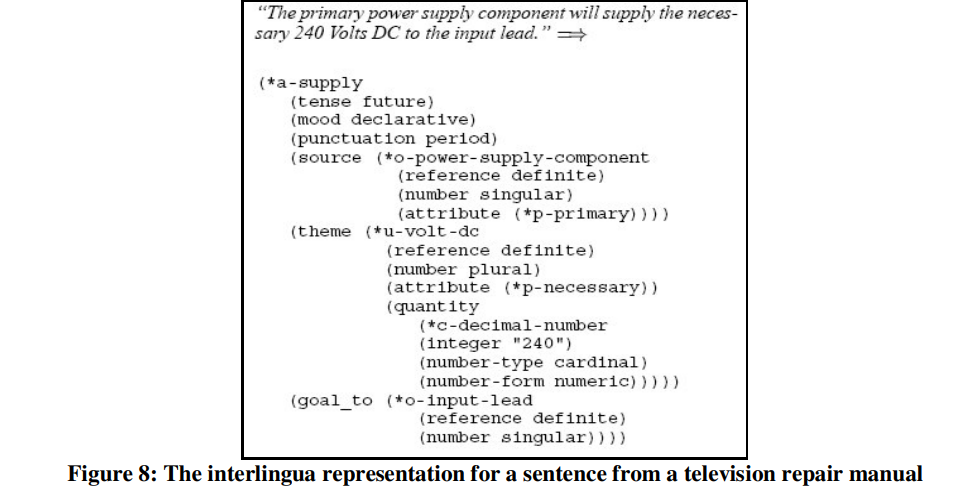


学过编译原理、写过Compiler的同学容易看出这一方法（至少在encode阶段）与编程语言的相似之处，这里的Interlingua或许可以类比汇编/机器语言。

尽管有这些（理论上）的优点，但基于中间语言的机器翻译实际上却是三种经典方法中用得最少的。这主要是因为以下几方面难题：首先，自然语言本身是有很多歧义的，因此分析模块经常需要从多种可能的中间语言表征选择一种，对此有的模型（如DLT）维护了Lexical Knowledge Bank（LKB），它实际上是一个数据库，用于指出哪个单词在给定的上下文中最有可能出现（我们可以看到在这里模型已经不完全是rule-based了）。其次，尽管中间语言的设计出发点是对人和机器“思维方式”的折衷，即，它既对人类是可读（从而可编写和修改的），又对机器是可处理（无歧义）的，但在实际构造过程中，它通常较为晦涩，可读性很差而且修改起来非常困难。比如麻省理工的UNITRAN模型构造的中间语言表征Lexical Conceptual Structure（LCS）将一个简单的英语句子表示如下：



卡内基梅隆大学的KANT（一度用于翻译技术文件）作为基于中间语言的为数不多的商用机器翻译模型。它被设计为只翻译源语言的一个特定领域子集（技术文档，如电子产品使用说明书），严格限制处理的词汇和句法结构。在足够狭窄的领域里，词汇的含义比较固定（如clean在说明书中通常表示“清理”而非别的），句型也相对单一。然而这样并没有解决歧义问题，而是回避了它。



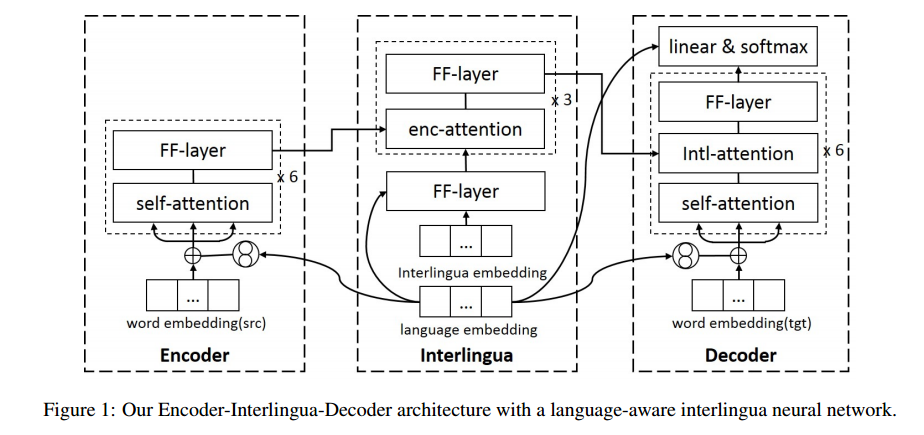
除此之外还有其他一些质疑，如Açıkgöz和Sert认为把自然语言压缩成一套通用符号“否认了语言的本质和文化价值”。任何中间语言，不管设计得多么精妙，它能够表示的含义仍然只是自然语言一个小的子集（他们仍然固守Halliday的观点，既code和message是不可分割的），当然，在这里不展开。

基于中间语言的机器翻译存在规则复杂、语言特征损失和歧义等问题，因此多数研究都停留在实验室阶段，极少投入使用，并且随着神经网络机器翻译的出现逐渐淡出人们的视线。在这里我想顺便提到另一个问题，即人类对语言和智能的理解是不断变化的。语言与智能在进化中是紧密联系在一起的，通常被认为是人类与动物、与机器的区别所在。John Searle的Chinese Room Argument和图灵测试背后的逻辑也有相通之处：理解和掌握一门语言对造出能正确表意的句子是必须的吗？机器一定要拥有“智能”才能在交流中表现得足够像人吗？

随着词向量、语义空间（毕竟，万物皆可embedding嘛）的诞生和流行，对使用一种人造中间语言来表示文本含义的追求也许可以换一个思路，即，尽管我们常常认为（刻板印象！）深度学习是一个黑箱，但也许其实矩阵、向量、参数这些东西本身就是机器掌握和记忆的知识，不能因为人类读不懂就否定它。与其费力地人为设计一种机器能处理的无歧义的中间表征，并辅以复杂的分析和生成规则，为什么不让机器自己构建它自己能“读懂”的中间语言呢？

多语言神经机器翻译（Multilingual neural machine translation，NMT）和基于中间语言的机器翻译具有共同的目标，即追求经济性。后者希望能够制定更少的规则，前者则节约了模型的资源，且通过跨语言共享语言学信息能够适配一些低资源（缺乏大规模平行语料）的场景。然而，传统的多语言模型，就像基于中间语言的模型一样，很难捕捉到不同语言的多样性和特殊性，导致与经过充分训练的单个模型相比性能较差。对此，阿里巴巴实验室在编码器-解码器架构中加入能针对不同语言特点进行适应的中间语言结构。模型能够从不同语言的语义空间中学习与语言无关的表示，同时仍然允许针对特定语言的专门化。

具体来说，模型引入language embedding来表示每种语言的独特特征，以及interlingua embedding来捕获跨语言的共同语义。两个embedding用于增强自注意机制，将encoder表示转换到共享语义空间。为了尽量减少变换过程中的信息损失和保持语义一致性，训练目标引入了重构损失和语义一致性损失。



这是一个encoder-interlingua-decoder架构。其中encoder和decoder与一般的自注意Transformer相同，只是在position embedding中进行了修改。interlingua是跨语言共享的，但有基于特定语言的embedding作为输入。interlingua模块由多个相同的层组成，每一层都有一个多头注意力子层和一个前馈子层。在这个模型中，通过整合共享的interlingua embedding，来自相同和不同语言（不同语义子空间）的不同句子的相同语义成分应该映射到共享语义空间里相同的位置。

* Language-aware Positional Embedding

考虑语言之间的结构差异，每种语言都应该有其positional embedding。模型通过为Transformer中的原始正弦、余弦函数提供与具体语言相关的偏移量，从而为encoder和decoder提供每种语言独立的positional embedding。

Training Objective

1. Translation objective（这部分与其他神经网络翻译模型相同）

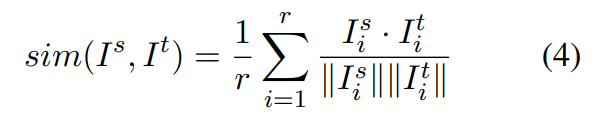
双语NMT模型通常采用交叉熵损失作为训练目标，记为Ls2t，表示从源语言到目标语言，同时引入另一个Lt2s，进行从目标语言到源语言的翻译。

1. Reconstruction objective（这部分可以类比“回译”结果与原文对比）

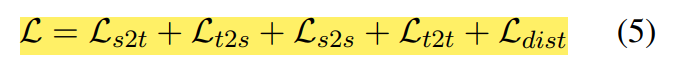
X’ = Decoder(Interlingua(Encoder(X)))表示X的reconstruction，使用X’和X之间的交叉熵作为其损失，并用Ls2s表示源语言，Lt2t表示目标语言。

1. Semantic consistency objective（这部分在zero-shot中有类似应用）

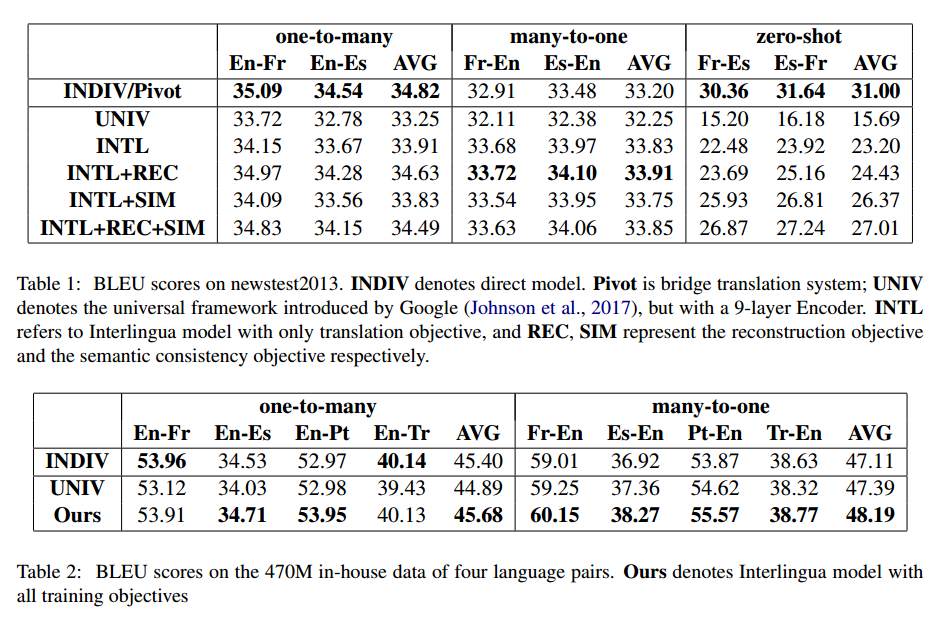
余弦相似度/距离（简单但有效的方法！）来衡量一致性（即用不同语言表示的同一个句子应该被映射到语义空间的同一位置）。Ldist = 1 − sim(Is, It)。



最后得到的目标函数就是：



这种language-aware“中间语言”多语神经机器翻译模型在许多语言对上的测试结果都接近精心训练的双语模型，而且显著优于通用多语言模型。



虽然embedding和语义空间的向量和传统对“语言”的认知相去甚远，但将它称为一种“中间语言”无疑不但富有智慧和创造力，而且是形象和贴切的。就像我们之前谈到的，随着人工智能、机器翻译和种种语言模型的发展，人类对语言和智能，“意义”和“理解”的定义和认识或许也应该与时俱进。

主要参考文献

Alansary, Sameh: Interlingua-based Machine Translation Systems: UNL versus Other Interlinguas. *The Egyptian Journal of Language Engineering*. 2014.

Halliday, Μ. Α. Κ. : Linguistics and Machine Translation. *STUF - Language Typology and Universals* 15. April 1, 1962.

Zhu, Changfeng et al: Language-aware Interlingua for Multilingual Neural Machine Translation, Machine Intelligence Technology Lab, Alibaba Group. ACL 2020.